TEADAL

# D2.2 Pilot cases' intermediate description and initial architecture of the platform

Revision: v.1.0

| | |
|---|---|
| **Work package** | WP 2 |
| **Task** | Task 2.1, 2.2, 2.3 |
| **Due date** | 30/11/2023 |
| **Submission date** | 30/11/2023 |
| **Deliverable lead** | Cybernetica (CYB) |
| **Version** | 1.0 |
| **Authors** | Eduardo Brito, Andre Ostrak, Pille Pullonen-Raudvere (Cybernetica) |
| **Reviewers** | Andrea Falconi (MARTEL), Pierluigi Plebani (POLIMI) |
| **Abstract** | This deliverable provides a second iteration of pilot use-case requirements, adding the shared financial data governance pilot. It details data generation processes for the pilot cases. It describes the first iteration of the general architecture of TEADAL. |
| **Keywords** | TEADAL, synthetic data, architecture |



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

## Document Revision History

| Version | Date | Description of change | List of contributor(s) |
|---|---|---|---|
| V0.1 | 31/10/2023 | 1st version of document for the advisory board | Ilaria Baroni, Alessio Carenini (*Cefriel*)<br>Eduardo Brito, Andre Ostrak, Pille Pullonen-Raudvere (*Cybernetica*)<br>Andrea Falconi (*Martel Innovate*)<br>Praveen Kumar Donta, Victor Casamayor Pujol, Boris Sedlak (*TU Wien*)<br>Jorge Catarino (*Ubiwhere*) |
| V0.2 | 20/11/2023 | 2nd version of document for internal review | Ilaria Baroni, Alessio Carenini (*Cefriel*)<br>Eduardo Brito, Andre Ostrak, Pille Pullonen-Raudvere (*Cybernetica*)<br>Andrea Falconi (*Martel Innovate*)<br>Praveen Kumar Donta, Victor Casamayor Pujol, Boris Sedlak (*TU Wien*)<br>Jorge Catarino (*Ubiwhere*) |
| V1.0 | 29/11/2023 | Integration of reviewers' suggestions | Eduardo Brito, Andre Ostrak, Pille Pullonen-Raudvere (*Cybernetica*),<br>Victor Casamayor Pujol (TU Wien) |

## Disclaimer

Funded by the European Union

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.

## 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.



Funded by the European Union

# EXECUTIVE SUMMARY

Trust, privacy, and energy efficiency have rightly become key concerns in today's data-driven era. While federated data sharing across organisation and computing boundaries fuels industry innovation and decision-making, often the involved parties engage in data sharing only after having gained sufficient trust that procedures and tools are in place. These must ensure that data are exchanged and processed solely according to agreed-upon workflows, privacy is strictly enforced, and computations are energy-efficient. In line with the TEADAL vision of federated smart data management, in a rapidly evolving data sharing landscape, this document investigates the challenges of trust, privacy, and energy efficiency in a federated setting and presents a federated data mesh architecture to address them.

The TEADAL pilot use cases provide insight into the business challenges inherent in federated data sharing and guide the design of the TEADAL architecture. Previously released TEADAL documents investigated these challenges in various domains, each with unique requirements: healthcare, mobility, viticulture, industry, and regional environmental development. The present document updates these use cases and gives a clear overview of the data generation processes for each one. Moreover, it introduces a new shared financial data governance use case, involving an industry partner and offering a comprehensive view of the scenario, data flows, analytics, and synthetic data generation.

The TEADAL architecture embraces both the service and data mesh paradigms, developing novel ways to address trust, privacy, and energy efficiency in federated data lakes. TEADAL extends the concept of data product to that of federated data product so that data can be shared according to the governance rules of a given federation. A catalogue allows consumers to discover federated data products and enter into a data-sharing agreement with a producer to consume a subset of the available federated data product. This leads to the concept of shared federated data product, which encapsulates the actual process of sharing data between a producer and consumer, according to a specific producer-consumer agreement. Data products are offered as RESTful services in a service mesh where proxies intercept service communication. Interception allows TEADAL to track the data product life-cycle and produce verifiable evidence, through blockchain technology, that only the intended parties exchange data according to an agreed-upon workflow. Additionally, an infrastructure-as-code approach to deployment allows TEADAL to track all the software involved in these interactions. Security and privacy enforcement also leverage message interception to allow or deny access to data products according to policies expressed in a business-oriented, high-level language. Finally, TEADAL introduces the concepts of gravity and friction to optimise mesh resource usage and placement along the computing continuum. Through the service mesh observability function, TEADAL can monitor and attempt to reduce data-sharing overhead among organisations (friction) as well as allocating computations based on data proximity versus processing power trade-offs (gravity).

It should be noted that the architecture documented here is the result of the first of three planned design iterations. As such, some parts of the architecture still require further development whereas other parts need refinement. The implementation of the pilot use cases will provide valuable feedback to adjust and steer the initial design. However, this first iteration of the architecture does provide the necessary foundation upon which the next iterations can build.



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

| | |
|---|---|
| **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** | General Transit Feed Specification |
| **GTFS-RT** | General Transit Feed Specification realtime |
| **IaC** | Infrastructure as code |
| **IAM** | Identity and access management |
| **jwt** | JSON Web Token |
| **KPI** | Key performance indicator |
| **KYC** | Know your customer |
| **MDM** | Master Data Management |
| **ML** | Machine learning |
| **NAP** | National access point |
| **OLAP** | Online analytical processing |
| **OLTP** | Online transaction processing |
| **OMOP** | Observational Medical Outcomes Partnership |
| **PDP** | Policy decision point |
| **PII** | Personally identifiable information |
| **RAP** | Regional access point |
| **SaaS** | Software as a service |



Funded by the European Union

| | |
|---|---|
| **SFDP** | Shared federated data product |



Funded by the European Union

# 1 INTRODUCTION

Data is a priceless resource in the digital age, driving innovation, decision-making, and progress across industries. The ability to collect, analyse, and share data has become pivotal, facilitating the development of smarter technologies, efficient processes, and better-informed societies. However, as the importance of data sharing continues to grow, so do the associated challenges and concerns. Individuals and organisations must have confidence that their data will be handled securely, ethically, and in a manner that respects privacy. Trust is built through transparency, accountability, and adherence to ethical data practices. This trust, in turn, enhances collaboration and unlocks the true potential of data sharing. Responsible data management aligns well with the environmental objectives of data minimisation, including more energy-efficient data storage, sharing, and consumption.

The TEADAL project aims to provide a toolset to allow data providers and consumers more trustworthy, privacy-aware, energy-efficient data usage in a federated setting. The project entails six pilot use-cases from different domains - healthcare, mobility, viticulture, industry, finance, and regional environmental development - each providing unique requirements that the proposed solution should fulfil. In most cases, these requirements, as-is and to-be scenarios, data descriptions, business process models, and data generation processes have already been detailed in deliverable D2.1 of the TEADAL project.

The current deliverable mainly consists of two parts. First, it will briefly describe the pilot cases, their problem statement, and goes into more detail on the data generation. The process of data generation has been different for each use case and has required thorough investigation to make sure the data fits the purposes of the project, providing the necessary data integrity while also making sure that no sensitive, private, or confidential data is leaked. The project uses real open data from both external and internal parties of the project, synthesised data from different partners, and real anonymised or non-sensitive data from the pilot partners.

An important outlier of the pilot use-cases is the shared financial data governance. This deliverable is filling in a section that was missing from deliverable D2.1 as the corresponding information was delayed. Chapter 7: USE CASE PILOT 5: SHARED FINANCIAL DATA GOVERNANCE describes the pilot use-case in detail, providing the as-is and to-be scenarios, the data flows and analytics performed, the musts and wants for the project, an overview of the processes in a business process modelling notation, and a description of the synthetic data generation.

The deliverable also gives the first iteration of the general architecture for the TEADAL toolset and framework, including a high-level overview, the design choices and how they were identified from the first iteration of the pilot requirements elicitation. It describes the components and how they fit together into a cohesive whole, providing a view of the creation, discovery, usage, and the discontinuation of federated and shared federated data products, which are concepts introduced in the project. It explains how the proposed architecture leads to a more secure, trustworthy, privacy-preserving, and energy-efficient data exchange in a federated setting. In addition to a high-level conceptual view, the deliverable goes into more detail regarding the current iteration of the deployment of TEADAL tools, describing the software components used and how it fits together with the current architectural design. This is to make sure that the deployment is consistent and follows the defined architecture, in the current iteration as well as all the subsequent ones.

The main chapters of the deliverables can be divided into two distinct parts. First, chapters 3 to 8 focus on the pilots.



Funded by the European Union

Chapter 3, USE CASE PILOT #1: EVIDENCE-BASED MEDICINE, gives a brief overview of the pilot use-case, goes into more detail on the anonymisation/synthetic data generation process by the pilot partner MARINA SALUD and explains why the generated dataset is sufficient for the project.

Chapter 4, USE CASE PILOT #2: MOBILITY, gives a brief overview of the pilot use-case, explains how the data from the pilot partner AMTS was gathered, how synthetic data was generated, and how open data was collected. It will also go into detail about the data ingestion.

Chapter 5, USE CASE PILOT #3: SMART VITICULTURE, gives a brief overview of the pilot use-case, explains how the data is gathered by the pilot partner TERRAVIEW, and how the data ingestion is done. It details some changes for the data analytics of the pilot use-case from deliverable D2.1.

Chapter 6, USE CASE PILOT #4: INDUSTRY 4.0, gives a brief overview of the pilot use-case, and explains how the data is gathered by the pilot partner ERT. It details some changes to the pilot use-case from deliverable D2.1.

Chapter 7, USE CASE PILOT #5: SHARED FINANCIAL DATA GOVERNANCE, describes in detail the pilot requirements, the to-be and as-is scenarios, the synthetic data generation, process models and data flows.

Chapter 8, USE CASE PILOT #6: REGIONAL PLANNING FOR ENVIRONMENTAL SUSTAINABILITY, gives a brief overview of the pilot use-case, explains how the data was gathered and synthesised. It will also explain the data ingestion for the pilot.

Chapter 9 describes the first iteration of the TEADAL general architecture, providing the conceptual view and explaining the design choices, as well as the choices currently made for the deployment of the baseline. This includes the design and the corresponding tools of the TEADAL node.

In addition, the deliverable describes the methodology in Chapter 2, and finally gives conclusions and future work in Chapter 10.



Funded by the European Union

# 2 METHODOLOGY

TEADAL requirements collections are designed with an iterative approach, where pilots' requirements are refined periodically, thanks to the feedback coming from the technical implementation. This document represents the second iteration of the process, describing the pilots and specifying their requirements. While the focus of the first iteration was on providing high-level requirements (see D2.1 for more information), the purpose of these documents is to update and provide further details for the use cases requirements, the data and its synthetic generation processes.

Following the methodology, a second round of workshops was implemented between June and July 2023, one for each pilot. Each workshop took one hour and involved both technical and use cases partners. They were focused on components update (datasets, data products, policies, etc...) and requirements update (any changes with respect to D2.1). After the meetings, no further information was raised with respect to the previous deliverable for the five pilots described in D2.1. In addition, the requirements for the "Shared financial data governance" pilot (due for deliverable D2.1, but which were not included because of administrative difficulties regarding the pilot) are described in this document.

In cooperation with the pilots, the available datasets were discovered, detailed, and analysed. According to the requirements, the respective data was gathered, or synthetic data was eventually generated, following the schemas and formats reported by the pilots. Multiple individual meetings took place in order to discuss the aspects of such activities and to reach agreement with the pilots, regarding the modelling of their business processes and the work to be done in order to produce the data, in both quality and quantity, suitable for the rest of the project's plans. In some cases, the pilot partners took the responsibility of identifying, gathering, generating, and auditing their data, and, in other cases, these tasks were delegated to Cybernetica, which led, supervised, and engineered the required processes. These activities benefited from multiple exchanges of feedback and information, both synchronous and asynchronous, between Cybernetica and the pilot partners, which resulted in the development of the data ingestion mechanisms detailed in this deliverable. In more broader discussions, involving the technical partners of the consortium, the pilots' goals for data consumption and analysis were further dissected, derived from early plans documented in the deliverable D2.1, and adjusted to the evolving federated data product definition, as well as estimations of the resources needed to host the project's infrastructure. It is worth mentioning, however, that all the pilots present different levels of maturity, and the work done up to this stage is likely to continue to mature, as better and more accurate insights are gathered. The pilots' specific chapters of this deliverable detail these differences, as well as the multiple approaches taken to reach the current state of data description, generation, and ingestion.

TEADAL's architecture progress has been achieved through an iterative process of joint technical meetings and architecture prototyping. In brief, the joint technical meetings have defined the architecture components and interactions, which have then been prototyped to specific technical tools. The prototype has been reviewed in the joint technical meetings, which, as an outcome, has defined new components and interactions, or improved the ones already present. The joint meetings dedicated to the development of architecture are as follows:

- Architecture preliminary discussions: 7&15 December 2022 (2 online sessions)
- TEADAL technical joint discussions: 9 & 11 January 2023 (3 online sessions)
- Architecture components brainstorming: 13 January 2023 (1 online session)
- TEADAL's architecture beyond Data Mesh: 27 February 2023 (1 online session)
- 2nd TEADAL's GA: 15-16 March 2023 (Denia - Spain)



Funded by the European Union

- TEADAL architecture vision with Data Mesh and Serverless exchange: 22 & 31 March and 3 & 5 April 2023 (4 online sessions)
- Specific architectural technical meetings (in person): 26 & 27 July 2023 (Milan - Italy)
- Architecture development technical meetings: September 2023 (4 online sessions)
- 3rd TEADAL GA: 10-11 October 2023 (Online)

The work on architectural design is ongoing and converging fast with the pilot use-cases.



Funded by the European Union

# 3 USE CASE PILOT #1: EVIDENCE-BASED MEDICINE

## 3.1 PILOT OVERVIEW

The aim of the evidence-based medicine pilot is to improve the current status of data analytics in healthcare, easing the process of sharing medical data. The pilot will focus on data privacy constraints that are the main handicap in healthcare analytics. Due to such constraints, studies are mostly performed over anonymised data. Given its complexity, previous initiatives at the EU level haven't addressed this issue. In this pilot, MARINA[1] will simulate federated data sharing among health organisations, and TEADAL will implement mechanisms to address data privacy constraints. These should happen both at the organisational level, by allowing the establishment of trustworthiness between organisations, if requirements for data access are met, and at the individual level, by providing tools to automatically enforce, or restrict, the usage of data only from individuals that consented to participate in the medical studies. A thorough description of the pilot, its stakeholders, goals, and project requirements can be found in deliverable D2.1, under the "USE CASE PILOT #1: EVIDENCE-BASED MEDICINE" chapter. The main update to the pilot's requirement is the use of multi-party computation for the secure aggregation of the number of patients relevant to a study promoter. This privacy-enhancing use case is more thoroughly explained in deliverable D5.1.

## 3.2 DATA DESCRIPTION

The existing data, its characteristics, volume, and accessibility was described in detail in D2.1, under the section 3 of the "USE CASE PILOT #1: EVIDENCE-BASED MEDICINE" chapter. Data interoperability is the cornerstone for collaborative (federated) analytics in healthcare. To address it, the pilot has selected the OMOP[2] (Observational Medical Outcomes Partnership) standard as data model. It is a recognized international standard that makes it possible for data provided by one organisation to be understood and used by another, enabling as well the provision of data by multiple organisations to conform to larger datasets, which is needed to generate medical evidence. A subset of the OMOP data model has been selected, containing seven different datasets: Person, Procedures, Conditions, Observations, Visits, Drugs (medications) and Measurements. This subset includes the most relevant items in an Electronic Health Record, supportive of a wide variety of clinical studies.

## 3.3 SYNTHETIC DATA GENERATION

In this pilot, only synthetic data is used, as medical data is of the most sensitive kind, from the perspective of data privacy. However, the synthetic data has not been generated from scratch, otherwise it could have resulted in data that may not represent the clinical reality, in terms of syntax, and even semantics. The generation process has been performed only by personnel from MARINA SALUD, through three different phases:

1. Obtaining primary pseudonymised data: The source data has been obtained from the corporate data warehouse of MARIAN SALUD, which is a backup SQL database that is updated daily and stores all the new records generated from patients. This backup data warehouse exists to reduce the load and impact of directly accessing the healthcare production database to perform analytics, as it could induce performance

[1] https://www.marinasalud.es/
[2] https://www.ohdsi.org/data-standardization/



Funded by the European Union

issues in a system that has to be up and stable at all times. An ETL process owned by the hospital has produced a reduced OMOP formatted version of the healthcare records, from approximately 100k patients. In this OMOP dataset, the PERSON object has a person_id that serves as a link to the rest of the objects (drugs, visits, procedures, etc...). This PERSON object has been modified to only store basic parameters, such as age and sex, and hide or exclude other demographics that could serve to identify the subjects behind the data.

2. Anonymising the data: The result of the first step has produced a pseudonymised dataset in which only the person_id could be used to identify the data subjects. In this second step, the person_id has been replaced by a random number. Since the original person_id has not been kept, there is no way to identify the real data owner, even from inside MARINA SALUD. Thus, this is an anonymised dataset.
3. Synthesising data: MARINA has decided to go one step further in the anonymisation process, because, at this point, the persons could not be identified, but the subset of clinical history is still tied to a real individual, which perhaps may turn, in the future, into an identifiable datapoint. By this reason, the person_id attributes present in the surrounding objects (measurements, observations, procedures, conditions, drugs, and visits) have been replaced by randomly selected person_ids from the PERSON dataset. That way, the content of these ancillary objects remains real, but the clinical record of these anonymised patients has been changed in a way that there is no real data, and therefore can be defined as synthetic.

The result is seven datasets, in a standard format that can be used to assess analytical processes, producing results that may not be used to generate clinical evidence, which is anyways out of the scope of the project. The integrity of the synthetic data is enough to verify the correctness of the processes and analytics run during the project. The queries and statistical analysis will be deterministic and can be run on the synthetic datasets outside of TEADAL. Therefore, to make sure that all processes are carried out correctly within the TEADAL framework, it is possible to run the same processes on the synthetic datasets, without using TEADAL tools, and verify that the outputs are the same in both cases. The data currently resides in the MARINA SALUD infrastructure, which is sufficient for the project, as they are the infrastructure providers for this specific use-case. The data is available to the consortium per request.



Funded by the European Union

# 4 USE CASE PILOT #2: MOBILITY

## 4.1 PILOT OVERVIEW

The mobility pilot, based on TEADAL technologies, is set to demonstrate data sharing between four Italian entities: regional transport operator (AMTS), national transport operator (Trenitalia), Regional Access Point (RAP), and National Access Point (NAP). Since regional data collection initiatives in urban areas are limited due to disparate cross-border cooperation, Italy has delegated transport data collection to regions, creating a three-level system, where a Regional Access Point (RAP) collects data from transport operators and infrastructure managers and makes it available to the National Access Point. 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. However, since the writing of deliverable D2.1, this pilot case has undergone a few changes, which we will detail here. The available data has seen the addition of OpenStreetMap[3] and DATEX II[4] parking and weather data.

## 4.2 DATA DESCRIPTION

The existing AMTS data, its characteristics, volume, and accessibility was described in detail in D2.1, under the section 3 of the "USE CASE PILOT #2: MOBILITY" chapter. The data follows the described GTFS and GTFS-RT formats, encapsulating, among other information, relevant data points regarding the public transport scheduling, locations, real-time positioning, and delays from the municipality of Catania, Italy. The GTFS dataset is updated twice a year, and the GTFS-RT at every minute. The individual GTFS records take less than 1GB, and the individual GTFS-RT records take a few KBs in size. Other transport operator data follows, as well, the GTFS format for static transport datasets, with a few MBs in size and no officially specified timeframe for updates. OpenStreetMap data of Catania's territory was also collected. Additional parking and weather data, useful for the pilot's future infrastructural and analytics plans, follows the DATEX II format.

## 4.3 SYNTHETIC DATA GENERATION

AMTS provides a means, through a web server[5], for downloading the most up to date GTFS and GTFS-RT data of the public transport in the region of Catania. The repository is, by default, only accessible in Italy. However, AMTS provided access to the consortium as necessary. The GTFS dataset consists of static data, including the location of bus stops, routes, trip coordinates, and schedules. The GTFS-RT data provides the real-time tracking of the buses, including current location via a GPS device. These datasets are served in separate files, containing no personal, sensitive, or private data. The web repository does not guarantee data persistence, nor does it allow for querying old data. Both formats have different update intervals. Therefore, this pilot does not need to generate synthetic data, but instead to adopt a strategy to capture and retain the records that are uploaded at defined time intervals, for the different data formats. The result is a temporal snapshot of the real data.

[3] https://www.openstreetmap.org/
[4] https://datex2.eu/
[5] https://82.191.238.171



Funded by the European Union

The strategy relied on Apache Airflow[6], for the definition, scheduling, and maintenance of workflows, to download and persist the updated records. Two Airflow Directed Acyclic Graph (DAGs)[7] were created for, respectively and separately, downloading the GTFS and GTFS-RT updated records[8]. The first has a scheduled interval of 6 months, and the second operates at every minute. Both tasks perform a GET request to the web server, and save the binary response in the local filesystem, generating a uniquely timestamped identifier for the files to be saved.

The other data, namely Trenitalia[9] and OpenStreetMap[10], was manually collected from the respective websites. DATEX II data was generated from an existing schema[11]. Since its use is not tied to the pilot's project KPIs, its existence serves only the purpose of representing data that can be federated across the NAP and RAP boundaries. All the other data can and may be used for calculating the pilot's analytical goals at the end of the project.

[6] https://airflow.apache.org/
[7] https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html
[8] https://gitlab.teadal.ubiwhere.com/teadal-tech/airflow-mobility
[9] https://www.trenitalia.com/it.html
[10] https://www.openstreetmap.org/
[11] https://www.datex2.eu/schema/1_0/1_0/DATEXIISchema_1_0_1_0.xsd



Funded by the European Union

# 5 USE CASE PILOT #3: SMART VITICULTURE

## 5.1 PILOT OVERVIEW

In the last 5 - 10 years vineyard operators and wine makers have been facing increasing operational difficulties, due to climate change, to more strict compliance regulations and to a change in the expectations of modern consumers. In 2021, Terraview launched a SaaS platform, called TerraviewOS[12], that offers vineyard operators a system for better managing their assets. TerraviewOS helps vineyard operators in optimising decision-making processes by providing data driven foresights and in giving them adequate and timely warnings to deal with adverse conditions like pests, diseases and weather-related challenges. This product has been augmented by the development of Aquaview[13], a service that enables any customer with agricultural land to understand their water use through virtual water moisture maps. Each map is a set of contiguous water moisture probes that have an accuracy of +/- 1.5% of hardware-based control sensors.

TEADAL will help to develop a solution for enabling the data sharing between different vineyards, especially those placed next to each other, with the goal of quickly monitoring the changes that can have an impact on nearby vineyards (e.g., disease alert, weather data, specific spray diaries via traceability, …), but specifically water moisture profiles (surface level, and depth soil moisture).

## 5.2 DATA DESCRIPTION

The existing data, its characteristics, volume, and accessibility was described in detail in D2.1, under the section 4 of the "USE CASE PILOT #3: SMART VITICULTURE" chapter. The input data required for the pilot does not differ from what was reported in deliverable D2.1. What has changed is a bigger focus of the analytics carried out on the input data. The result of the analytics process is what the pilot defines as "Soil Moisture Maps". These are GeoJSON[14] files that annotate geographically set polygons with the data related to each polygon's soil moisture data. These GeoJSON outputs are ultimately the basis of the Federated Data Product (FDP), as known in Teadal architectural terminology. Should these FDP be made available (with a respective policy) for other customers by the owing customer, or Terraview, then they become the shared Federated Data Product (SFDP). A minor update to the information contained in D2.1 is that Terraview is no longer hosted on Google Cloud but Microsoft Azure.

## 5.3 DATA GENERATION

In the Smart Viticulture Pilot, the generation of data is not synthetic. The pilot uses public sources of data (satellite constellations of Sentinel[15] and Landsat[16]), along with the customer's location, expressed as an Area of Interest (AOI) in GeoJSON. The pilot is able to process and analyse the AOI with respect to soil moisture. Should a customer require depth models for their soil moisture map, they need to supply a sample of this data, which itself

[12] https://www.terraview.co/
[13] https://aquaview.ch
[14] https://geojson.org
[15] https://sentinels.copernicus.eu/
[16] https://landsat.gsfc.nasa.gov/



Funded by the European Union

cannot be shared. The resulting FDP can be shared should the customer allow it and define the corresponding policy.

The data is gathered from Microsoft's Planetary Computer[17], which combines a multi-petabyte catalogue of global environmental data (specifically satellite imagery, in this case). The selected data is analysed and cached locally. Once data is analysed, the process starts by cleaning all the scenes from disturbances, and after this step the data is calibrated. The pilot then produces results and stores them into their storage infrastructure, which is then provided to customers through a user interface.

This process of ingesting, storing, analysing and finally offering is encoded in an Airflow Directed Acyclic Graph (DAG)[18] as shown below.

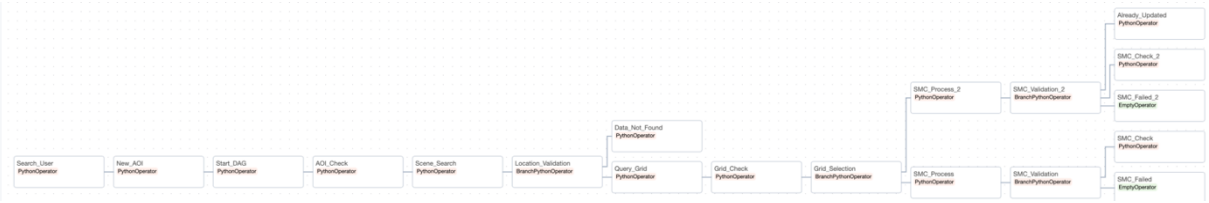


*FIGURE 1: SMART VITICULTURE'S AIRFLOW DAG.*

It is this DAG that will be deployed upon the pilot testbed as specified in deliverable D6.1.

[17] https://planetarycomputer.microsoft.com
[18] https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html



Funded by the European Union

# 6 USE CASE PILOT #4: INDUSTRY 4.0

## 6.1 PILOT OVERVIEW

The focus of the industry pilot is on the need for calculating a set of (key performance indicators) KPIs that are shared between two ERT Group[19] plants from different countries (Portugal and Czech Republic). As data is taken on a per-facility basis, the company-wide KPIs have to be calculated and accumulated according to the collective standard of the entire group. TEADAL aims to improve and automate the calculation of KPIs relevant for the management of the company (operational, commercial, quality, etc…), by providing tools to automate and optimise the technological impact of sharing data within ERT. 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.

## 6.2 DATA DESCRIPTION

The existing data, its characteristics, volume, and accessibility was described in detail in D2.1, under the section 3 of the "USE CASE PILOT #4: INDUSTRY 4.0" chapter. The data used in the analysis is owned by each plant of the ERT Group. Sales, logistics, production, quality, human resources, safety, and cash balance data is used to calculate the KPIs for a consolidated weekly report, from both plants. The data is updated frequently, however, only high level weekly updates are presented in the reports. A set of algebraic KPI formulas transform the raw data that feeds each plant's weekly report. The total volume of historical data is around 100GB, with a weekly volume of approximately 200MB. The pilot expects to calculate and present the required KPIs in a proper user interface, based on a set of defined access policies and permissions.

## 6.3 DATA GENERATION

ERT decided, in agreement with the TEADAL consortium, to provide all data as real data, from both industrial plants. Therefore, no synthetic data is needed. Both facilities generate information to be frequently inserted to a common cloud-based SQL database. It is composed of stocks, current balances, business partner balances, and financial balances. ERT has assured the TEADAL project partners that the data does not contain any personal or private data, and that it will be shared in accordance with existing agreements with their business partners. In infrastructural terms, the plants may share access to the database. The data curation and consumption may be decoupled from the ingestion and transformation processes. This architectural decision will be further shaped by the pilot, in accordance to TEADAL architectural plans. The aim of logically splitting the production and consumption systems is to ease the data normalisation process, from both plants, during ingestion, and to ensure the achievement of the pilot's goal of providing a centralised but access-controlled and policy-aware KPI visualisation component. Figure 2 shows ERT's updated business process model.

[19] https://www.ertgrupo.com/



Funded by the European Union

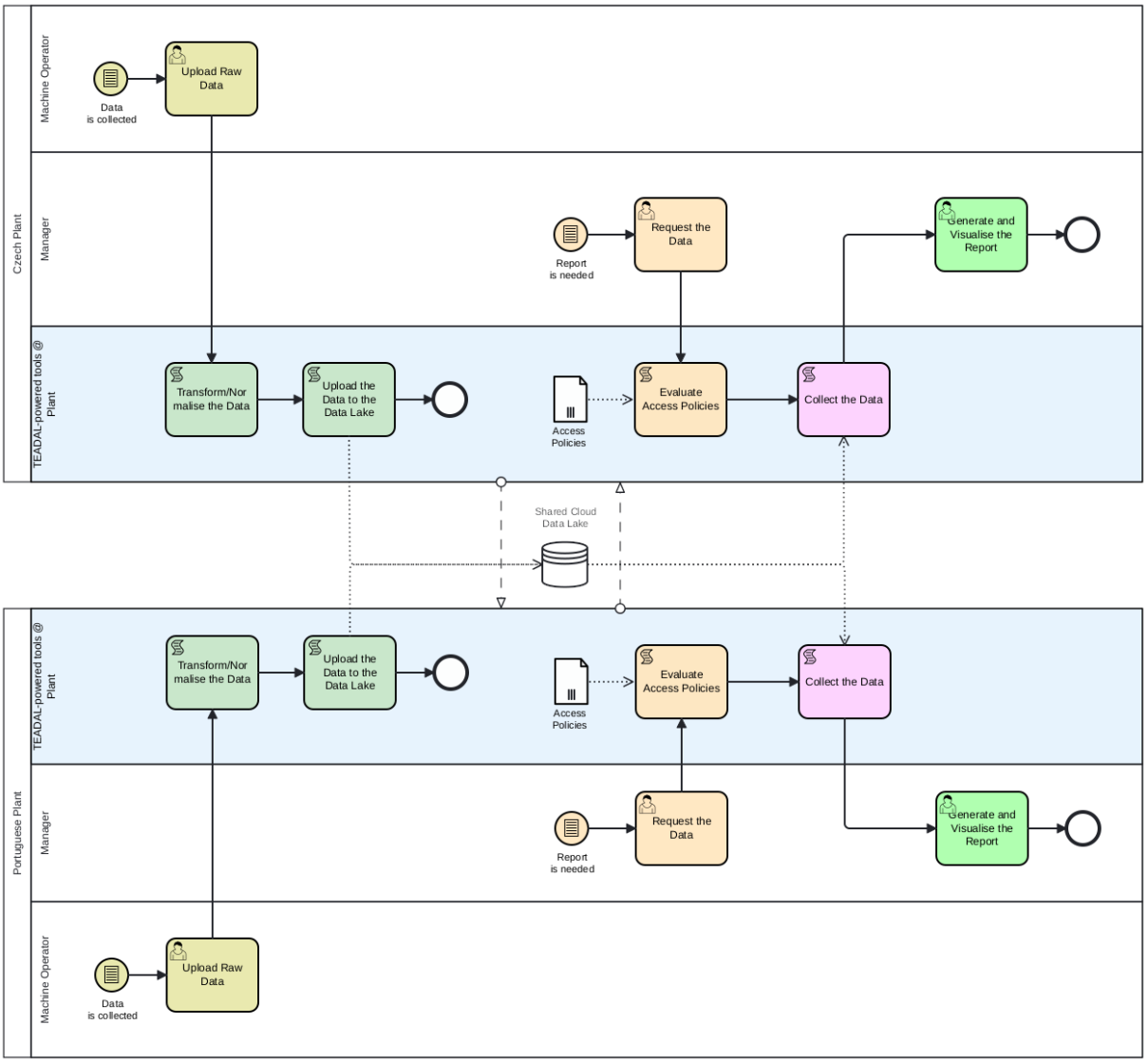


*FIGURE 2: TO-BE BPMN FOR THE INDUSTRY 4.0 USE CASE*



Funded by the European Union

# 7 USE CASE PILOT #5: SHARED FINANCIAL DATA GOVERNANCE

## 7.1 INTRODUCTION

Global financial institutions, such as international banks, are usually operating in multiple geographies. Operating on a global scale and in different domains pose number of challenges:

- Obligation to adhere to different regulatory, data governance and local sovereignty regulatory rules.
- Enabling efficient operations on a local and global scale, where data and the needed insights are dispersed across multi-cloud, hybrid environments and technology stacks.
- Too complex and often manual processes of identifying relevant data needed to support given activities.
- Ever-growing need to optimise, automate processes to ensure efficient and sustainable operation.

For compliant, efficient operation of the financial institution there are a number of activities that need to leverage data and insights across different domains and geographies in a governed and efficient way. An example of such activities are Know Your Customer(KYC) programs, Compliance Reporting, Fraud Detection, Operational Analytics.

With the many vital data rules, procedures, and regulations in place to protect the privacy of sensitive data, many businesses had to jump through numerous governance hoops, before using the data, to improve customer experience, provide better analytics results, or share data insights in another country. In this context, to ensure the consistent governance pattern, the challenge of ING[20] is more and more important to solve, as the governance can be extended from disparate data sources to data lakes in a multi-cloud environment, with smart data governance experience on discovering, accessing, provisioning, and sharing data. In particular, the pilot case will consider the local branches Enterprise/Global Domain Global KYC located in the Netherlands and two Local Domains being Turkey and Australia. Due to multiple regulatory institutes we need to cater for both global and local policies. In this context, TEADAL can provide a consistent, trusted, compliant, and automated foundation on enhancing data lake federation with smart and compliant data sharing in a context hyper-regulated as the financial domain. This is especially relevant when it comes to the complex enforcement of data residency policies, across the organisation, deployed in different business regions globally. The details of this pilot were not yet known when deliverable D2.1 was published, due to that, the following of this chapter goes into more depth about this pilot.

## 7.2 STAKEHOLDERS AND ACTORS

### 7.2.1 Stakeholders

The main stakeholders in the Shared Financial Data Governance use case are those who usually are integral in the process of providing and consuming data in the distributed data landscape.

[20] https://www.ing.com/



Funded by the European Union

Data Owner is the one who owns and understands the data, and who is responsible for granting permissions for use of the data. Data Owner together with Data Steward work on sufficiently describing data with metadata.

Data Protection Officer (DPO), together with compliance & legal experts, is responsible for providing advice and supervising compliance with and implementation of personal data protection.

Data Governance Officer helps define data governance policies.

Global KYC Business Analyst is responsible for creating a holistic customer view across the financial institution with compliant and predictive digital solutions for financial crimes.

Local KYC Business Analyst is responsible for make local business decisions/actions based on KYC transaction monitoring reports.

Analytics expert/Data Scientist is responsible for the creation of KYC models.

Global regulators (BCBS 239, ECB, etc) are defining regulations and guidelines that financial institutions need to meet to maintain banking licence. Such regulations usually span across multiple geographies.

Local regulators (DNB, Australia regulators) are defining regulations and guidelines that financial institutions within certain jurisdictions or geographies.

Enterprise regulators are internal regulators and officers within the enterprise/organisation. They could define additional rules and measures internally applicable to the enterprise.

Depending on personas and purpose, some stakeholders can see all the data and others have only a partial view. For instance, a global KYC (know your customer) Analyst/Manager may see the full dataset, while an Australian (AU) KYC Analyst/Manager is able to see part of the report pertaining to the AU customers.

**Regulatory constraints between stakeholders**

There are a number of regulations that an enterprise in the financial industry needs to adhere to maintain their licences. Legal/regulatory givens and constraints shape the interaction and the possibilities for data sharing between the stakeholders. Next regulatory constraints apply to the Pilot 5 Shared Financial Data Governance:

- GDPR (principles and regulation around identifying and processing of personal data and the right to be forgotten).
- BCBS 239 on data quality and traceability.
- Enterprise internal policies - for instance, data must be described with proper metadata and include confidentiality classification (public, confidential, secret, etc…). Depending on the metadata and purpose of the data usage, appropriate data protection measures are required, for instance, masking/reduction, and other processes.
- Local/country policies - for instance, if the data of a Turkish customer is in a specific system, the Turkish government requires access to that system. This might imply that there are limitations of moving Turkish customer data outside Turkey.

#### 7.2.1.1 Antagonistic stakeholder and false information



Funded by the European Union

Depending on the purpose of the data usage and type of stakeholder there might be strict policies on what a specific stakeholder or persona can see and use. The pilot highlights a scenario without antagonistic stakeholders or actors. The common aim is to provide true information and insights while adhering to regulation constraint.

### 7.2.2 Actors

In this use case, "customers" can be considered actors rather than stakeholders.

| Involved entities/Features | Description |
|---|---|
| Stakeholders | Domain/Country:<br>• Data Owner<br>• Data Steward<br>• Data Protection Officer<br>• Local KYC Business Analyst<br>• IT Policies Administrator<br><br>Global Organization:<br>• Data Owner<br>• Data Steward<br>• Data Protection Officer<br>• Data Scientist<br>• Global KYC Business Analyst<br>• IT Policies Administrator<br><br>External and Internal Regulators<br>• Global regulators (BCBS, ECB...)<br>• Local regulators (DNB, Turkish regulators, Australia regulators)<br>• Enterprise regulators (internal regulators and officers within the bank) |
| Actors | Data Provider<br><br>Data Consumer/Processor |
| Regulatory constraints that specify the data sharing between the stakeholders | Global regulations (BCBS 239, GDPR, data quality and traceability)<br><br>Enterprise regulators (internal regulators and officers within the bank).<br><br>Local/country policies |
| Legal constructs that shape the interaction and the possibilities for data sharing between the stakeholders (e.g., NDAs) | |



Funded by the European Union

| | |
|---|---|
| Antagonistic stakeholder | Given scenario is without antagonistic stakeholders or actors. The common aim is to provide true information, insights and assure compliance of the financial institution to the regulations. |
| Accessibility policies for stakeholders | Depending on the type of data and its metadata, the purpose of the usage and data protection policies. |

*TABLE 1: STAKEHOLDERS AND ACTORS SUMMARY FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

## 7.3 AS-IS SCENARIO

The AS-IS scenario is represented by the "Know Your Customer (KYC)" use case, illustrated in Figure 3.

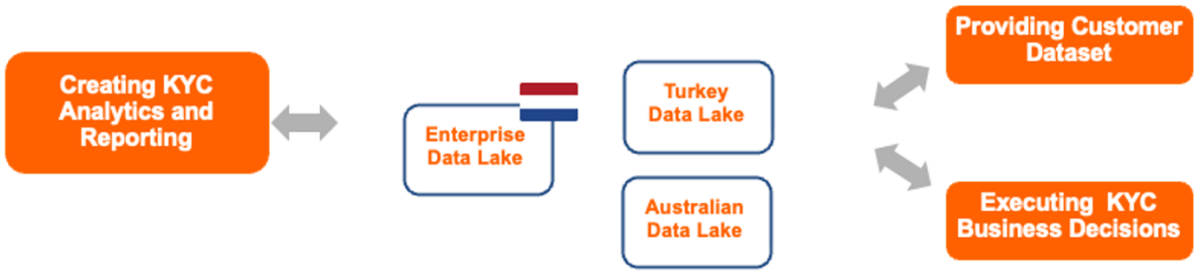


*FIGURE 3: SHARED FINANCIAL DATA GOVERNANCE AS-IS SCENARIO*

This scenario considers all the activities ensuring compliance with law, regulations, and policies to enable a financial institution to do continuous business with a customer within its risk appetite, from on-boarding to the exit of the relationship. The on-boarding requires consent gathering for KYC related activities. Analytics and reports are generated in every local branch. In addition, to create a holistic view of a customer, the data needs to be shared with the central global entity, who creates reports and alerts, which will in turn inform individual countries as required. While in some cases the data can be shared outside of a country's data lake, this is not possible in some countries due to local regulations. Furthermore, many of these processes are still not automated and optimised, and need more investigations and technical updates (orange boxes in the figure above).

Presently, data access is not monetised by ING but it represents a future objective. For analytical models, ad-hoc tools have been developed by ING and installed on premises. From a technical perspective, IBM Cloud Pak for Data[21] and Apache Kafka Event Bus[22] are used to implement data processing and to provide data feeds.

[21] https://www.ibm.com/products/cloud-pak-for-data
[22] https://kafka.apache.org/documentation/



Funded by the European Union

## 7.4 DATA DESCRIPTION

Currently, data is generated by processes, like "Master Data Management" (MDM) for customer data. The datasets may contain Personally identifiable information (PII), such as name, customer ID, customer bank account number, or address. Since this data fits under the category of sensitive personal data, with higher confidentiality rating, there is often a need for masking processes. For the development stage or purpose of generating models and reports, there is often a requirement to use anonymised, or synthetic data. All the data is owned by ING but some of it is bought from providers like Bloomberg, Reuters and other financial data providers. For Master Data we also check with Government Civil registries for correctness of the data (address and identification document validity). Data subjects are asked for consent by a paper-based questionnaire. In principle, this should be automated and registered in a catalogue.

### 7.4.1 Data Location and Format

The data is stored within ING, both in Europe and other geographies (Australia, Turkey). Local regulations and laws result in different restrictions on data movement/transfer to be put in place by companies acting in such geographies. The data used in the pilot is in four relational databases. KYC_CDD data consists of the customers data for Know Your Customer (KYC) and Customer Due Diligence (CDD) processes, while FATCA_DETAILS, G_ALERT_DETAIL, L_ALERT_DETAIL contain reports, as well as local and global alerts.

### 7.4.2 Data size and update frequency

The data can span from some TB (e.g., batch files) to a few KB in size (e.g., analytical results in Apache Kafka events). Size is, however, not the critical element of the financial use case. The creation and sharing of batch files is scheduled and has a fixed frequency. Copies of KYC databases are kept, and only changes are sent or processed, to save resources. Data confidentiality and integrity is guaranteed, as is its timeliness.

| Dataset | Personal/ Sensitive | Format | Location | Ownership | Size | Update Freq. |
|---|---|---|---|---|---|---|
| Global KYC | N | Relational | Netherlands | ING | 71 GB | Weekly/ Ad-hoc |
| Customers' data (Bank accounts) | Y | Relational | Romania | ING | ~30 GB | Weekly/ Ad-hoc |
| Customers' data (Real estate) | Y | Relational | Australia | ING | ~20 GB | Weekly/ Ad-hoc |

*TABLE 2: SUMMARY OF THE DATASETS AVAILABLE FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.4.3 Data sharing and federation

Since this pilot case is inherently distributed across different geographies and domains, data needs to be federated, as part of the overall KYC, CDD, and related activities. The data stored in local branches is expected to be shared with the global headquarters' operations team, and global reports and alerts generated from the holistic customer view are federated across the pilot's branches. This double-sided federation needs to be aware of local restrictions and regulations around data sharing, specific to some countries where the pilot operates.



Funded by the European Union

## 7.5 TO-BE SCENARIO

The TO-BE scenario represents the objective of the "Know Your Customer (KYC)" use case, described in the section 7.3 (AS-IS SCENARIO) and depicted in Figure 4.

In this use case ING is interested in creating customer's insights across different regions. The TEADAL project should support (i) creating analytics and reporting; (ii) providing customer datasets; (iii) executing business decisions.

This scenario considers all those activities ensuring compliance with law, regulations, and policies to enable an international financial institution to do continuous business with a customer within its risk appetite, from on-boarding to the exit of the relationship. The Know Your Customer(KYC) use case highlights the need for next aspects:

- Availability of relevant data for business-critical processes: *Data across multiple regions/environments.*
- Sufficient data description: *Regulation relevant description.*
- Necessity for governed shared data across domains/geographies: *Move or query data.*
- Data access controls.
- Enforcement of data policies: *Masking or reduction of data.*
- Intelligent data movement and protection: *Optimization based on costs and business policies.*

As depicted in Figure 4, customers' activity datasets are collected and sent to the headquarter where a team of Data Scientists and Business Analysts can provide a global report (Global Customer Profile Report). According to this report, an ING branch (e.g., in Australia) can interpret the behaviour of local customers (by the analysis of customers' activity datasets) and possibly identify suspicious customer profiles. Figure 5 illustrates the business process model for this use case.



Funded by the European Union

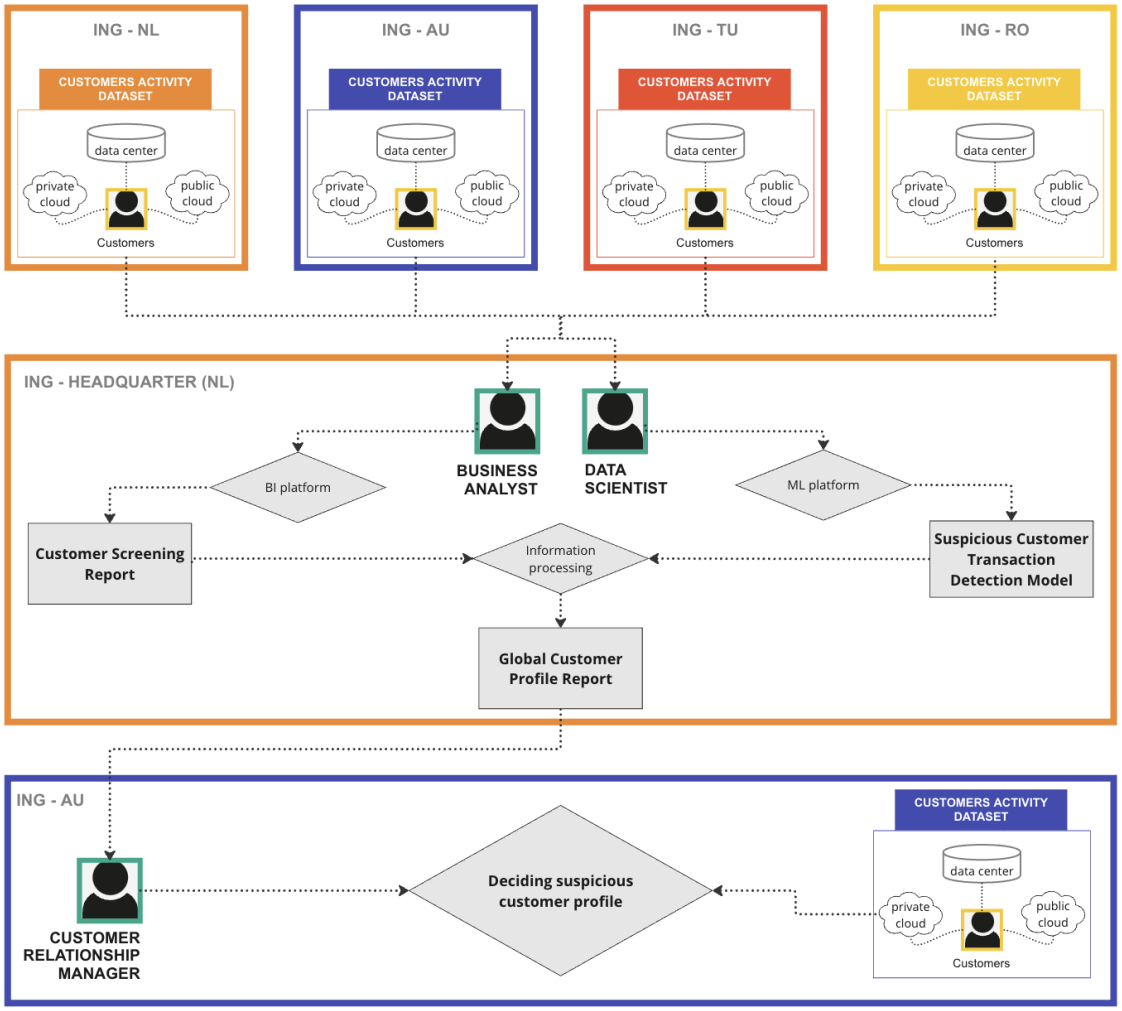


*FIGURE 4: TO-BE SCENARIO FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE: THE BUSINESS ANALYST AND THE DATA SCIENTIST (GLOBAL KYC TEAM) REPRESENT THE STAKEHOLDERS INTERNAL TO THE BANK. FURTHERMORE, THE CUSTOMER RELATIONSHIP MANAGER (AUSTRALIA KYC ANALYST/MANAGER) IS A STAKEHOLDERS INTERNAL TO THE BANK.*



Funded by the European Union

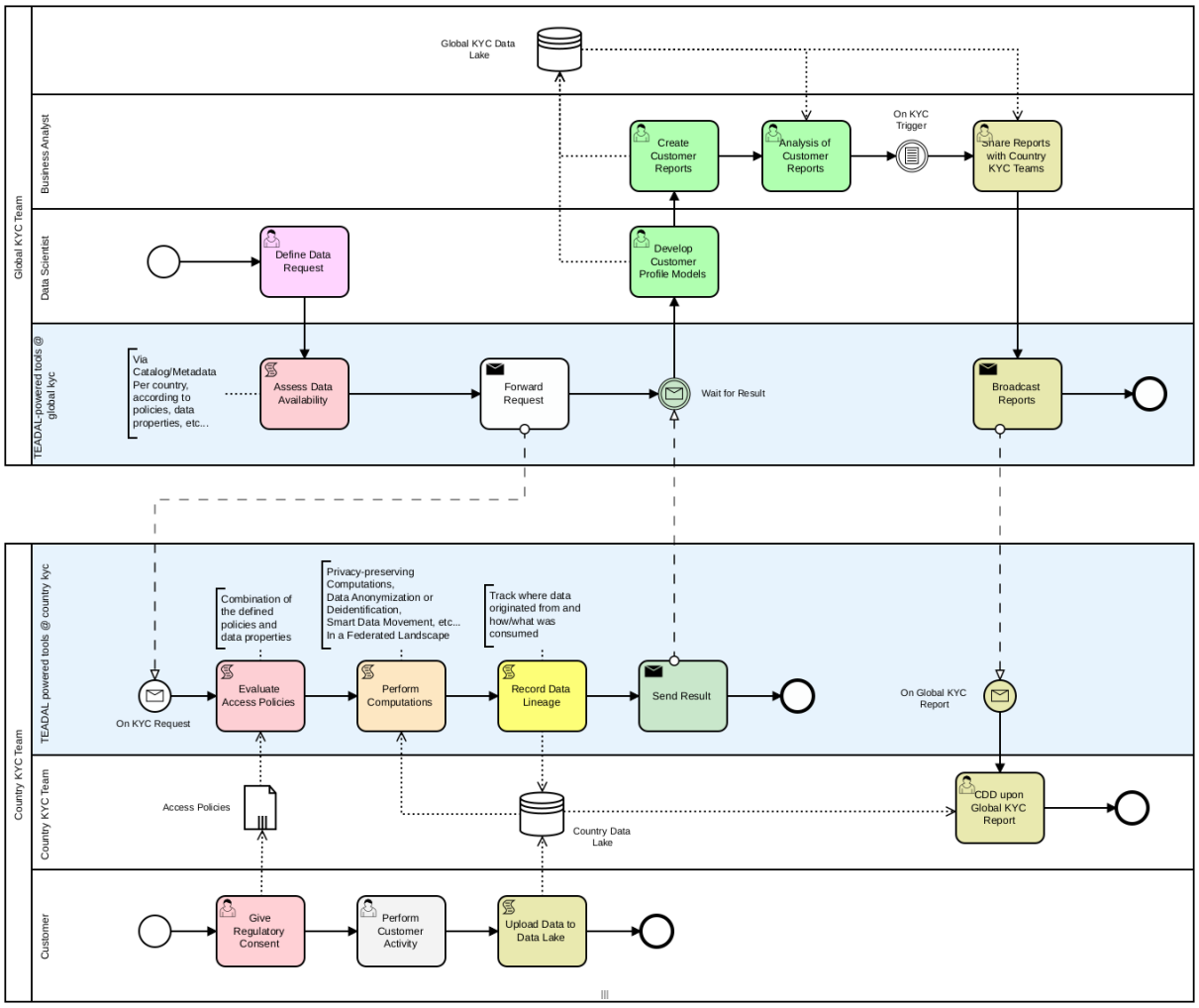


*FIGURE 5: TO-BE BPMN FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.5.1 Data analysis

Customer activity is collected, stored, and federated across the pilot's branches, in order to create a holistic view of the customer profile and activities. The main purpose is to allow for the creation of Customer Profile Models that feed global Customer Reports, shared with the local KYC teams. These reports allow for the local teams to perform CDD and KYC activities, in accordance with their local regulations and targeting goals for customer activity.

### 7.5.2 Data Usage and Privacy

Usage:

- Availability of relevant data for business-critical processes: Data across multiple regions/environments.
- Sufficient data description: Regulation relevant description.
- Necessity for governed shared data across domains/geographies: Move or query data.
- Data access controls: Move or query data.
- Enforcement of data policies: Masking or reduction of data.



Funded by the European Union

- Intelligent data movement and protection: Optimization based on costs and business policies.

Privacy:

Dataset may contain PII data such as name, customer ID, customer bank account number, address. Since this data is sensitive personal data with higher confidentiality rating often there is a need for masking. For the development stage/purpose of model/report there is often a requirement to use anonymised/synthetic data.

Online transaction processing (OLTP) and online analytical processing (OLAP) systems are both used. The data consumption is a series of batch ETL-jobs that are sequential, the individual countries deliver each and the central processing happens, after which the data set publishing back to the countries happens.

Also, model training (ML) and simulations are used. ING uses 5 years of history to develop the models, training of the productised models happens with 1 year of country data sets.

### 7.5.3 Data Location and federation

This pilot case's distribution across many domains and geographical areas requires the federation of data, as part of the broader KYC, CDD, and associated operations. The pilot's envisioned solution is to provide the global KYC team access to the data kept in local branches, and all of the pilot's branches will be federated to get global reports and alerts derived from the holistic customer perspective. The local laws and rules pertaining to data sharing that are unique to certain geographic locations where the pilot operates must be understood by this two-sided federation. The global KYC team is expected to operate in the Netherlands, while the Romanian and Australian branches represent the local KYC teams.

### 7.5.4 Data Models and synthetic data generation

Regarding the procedure for creating synthetic datasets, a masking tool is currently available. However, considering the high sensitivity of the data, the pilot will generate synthetic data based on a data schema provided by ING. The data consists of a relational database of four datasets, described in subsection 7.4.1, whose schema is common to all the pilot's local branches. The synthetic data will follow the provided schema with no regards made to the fidelity of the data, as the pilot seeks primarily to solve the data sharing and federation problem, in an infrastructural perspective. Real data can then replace the synthetic datasets once the pilot accomplishes its implementation goals.

## 7.6 TEADAL'S FEATURES

In the following table we summarise, according to the information currently available, the features of TEADAL which will be tested in this pilot. During the course of the project, new features could be investigated and tested.

| TEADAL's feature | Description |
|---|---|
| Data federation (Friction) | Multiple federated data lakes (customers activity data coming from several bank branches located in different countries). |
| Computation consumed in the continuum (Gravity) | Computation is not bound to be performed on the edge. |



Funded by the European Union

| | |
|---|---|
| Identity management | Multiple organisations with separate user bases, groups, and privileges. |
| GDPR compliance | Personal data is shared. |
| Data access tracking | Tracking access to data is not the primary focus of this pilot. |

*TABLE 3: TEADAL'S FEATURES SUMMARY FOR SHARED FINANCIAL DATA GOVERNANCE USE CASE*

## 7.7 REQUIREMENTS

In this chapter, a list of requirements is collected. They are divided into the five categories explained in the methodology chapter of D2.1 (general, privacy, architecture, data policy, data management), and they are listed by category and priority (must, should, could, won't).

### 7.7.1 General requirements

| Req. ID | Description |
|---|---|
| P5-Gen01 | Data should be tracked for knowing where data originated from and where it was consumed (Lineage) |
| P5-Gen02 | Computations should be executed on remote nodes and only the result should be moved on the central node |
| P5-Gen03 | The different environments (e.g., location) should be identified at the runtime/ execution level |
| P5-Gen04 | Both structured and unstructured datasets (also for streaming data) should be supported |
| P5-Gen05 | Effort required for Data Engineers to author and maintain data pipelines should be reduced |
| P5-Gen06 | Optimisation criteria to determine where computation and pre-processing takes place should be taken into consideration |
| P5-Gen07 | Data movement within a federated landscape should be considered |
| P5-Gen08 | Event-based transactions should be handled |
| P5-Gen09 | Energy-aware tools should be leveraged when moving data to faraway locations |
| P5-Gen10 | Local calculations and transformations required by national laws/rules should be dispatched and performed |
| P5-Gen11 | Means to notify applications that data or consent changed should be offered |

*TABLE 4: GENERAL REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.7.2 Privacy Requirements

| Req. ID | Description |
|---|---|
| P5-Privacy01 | Privacy should be preserved for computation tasks (for KYC model creation at least) |
| P5-Privacy02 | GDPR purpose consent flags should be handled independently |



Funded by the European Union

| | |
|---|---|
| P5-Privacy03 | GDPR consent gathering should be handled when building new analytics |

*TABLE 5: PRIVACY REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.7.3 Architecture requirements

| Req. ID | Description |
|---|---|
| P5-Arch01 | The facility to run per-data product purpose-based filtering should be provided |

*TABLE 6: ARCHITECTURE REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.7.4 Data Policy requirements

| Req. ID | Description |
|---|---|
| P5-Policy01 | Defined policies and data properties should be combined for applying respective the data product policies |
| P5-Policy02 | Data pipelines should be authored and compiled to automatically address hard and soft policies |
| P5-Policy03 | Monitoring data should be accessible on all data lakes (including an inventory of all available resources) |
| P5-Policy04 | Different implementations for the same pipeline depending on the policies should be attached |
| P5-Policy05 | Policies for time-to live of data sets should be defined |
| P5-Policy06 | A workflow for the additional consent should be available |

*TABLE 7: DATA POLICY REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.7.5 Data Management requirements

| Req. ID | Description |
|---|---|
| P5-Mgmt01 | Data will be accessed and processed from different geographical locations. So, data should be replicated for ensuring high availability. |
| P5-Mgmt02 | Data should be replicated for ensuring disaster recovery mechanisms |
| P5-Mgmt03 | Geographic location of data should be handled |
| P5-Mgmt04 | Metadata description of the data should be used as an input in the mesh for smart data movements |
| P5-Mgmt05 | Use the purpose of computation/job (development, production) to mask data if needed |
| P5-Mgmt06 | Use the purpose of computation/job to determine if data copy is needed |

*TABLE 8: DATA MANAGEMENT REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### 7.7.6 Requirements prioritisation

The collected requirements were prioritised and categorised in the following 4 classes:



Funded by the European Union

- *MUST*: mandatory functions for a baseline core system
- *SHOULD*: important functions, but not essential for the core system
- *COULD*: nice to have functions
- *WON'T*: functions that will not have right now or are out of scope of the project

## MUST

| Req. ID | Description |
|---|---|
| P5-Gen01 | Data should be tracked for knowing where data originated from and where it was consumed (Lineage) |
| P5-Privacy01 | Privacy should be preserved for computation tasks (for KYC model creation at least) |
| P5-Policy01 | Defined policies and data properties should be combined for applying respective the data product policies |
| P5-Mgmt01 | Data will be accessed and processed from different geographical locations. So, data should be replicated for ensuring high availability. |
| P5-Mgmt02 | Data should be replicated for ensuring disaster recovery mechanisms |
| P5-Mgmt03 | Geographic location of data should be handled |
| P5-Mgmt04 | Metadata description of the data should be used as an input in the mesh for smart data movements |
| P5-Mgmt05 | Use the purpose of computation/job (development, production) to mask data if needed |
| P5-Mgmt06 | Use the purpose of computation/job to determine if data copy is needed |

*TABLE 9: MUST REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

## SHOULD

| Req. ID | Description |
|---|---|
| P5-Gen02 | Computations should be executed on remote nodes and only the result should be moved on the central node |
| P5-Gen03 | The different environments (e.g., location) should be identified at the runtime/execution level |
| P5-Gen04 | Both structured and unstructured datasets (also for streaming data) should be supported |
| P5-Gen05 | Effort required for Data Engineers to author and maintain data pipelines should be reduced |
| P5-Gen06 | Optimisation criteria to determine where computation and pre-processing takes place should be taken into consideration |
| P5-Gen07 | Data movement within a federated landscape should be considered |
| P5-Arch01 | The facility to run per-data product purpose-based filtering should be provided |
| P5-Policy02 | Data pipelines should be authored and compiled to automatically address hard and soft policies |



Funded by the European Union

| | |
|---|---|
| P5-Policy03 | Monitoring data should be accessible on all data lakes (including an inventory of all available resources) |

*TABLE 10: SHOULD REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*

### COULD

| Req. ID | Description |
|---|---|
| P5-Gen08 | Event-based transactions should be handled |
| P5-Gen09 | Energy-aware tools should be leveraged when moving data to far locations |
| P5-Gen10 | Local calculations and transformations required by national laws/rules should be dispatched and performed |
| P5-Gen11 | Means to notify applications that data or consent changed should be offered |
| P5-Privacy02 | GDPR purpose consent flags should be handled independently |
| P5-Privacy03 | GDPR consent gathering should be handled when building new analytics |
| P5-Policy04 | Different implementations for the same pipeline depending on the policies should be attached |
| P5-Policy05 | Policies for time-to live of data sets should be defined |
| P5-Policy06 | A workflow for the additional consent should be available |
| P5-Mgmt07 | Storage for data type should be optimised |

*TABLE 11: COULD REQUIREMENTS LIST FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE.*

### WON'T

A central storage layer for historical data won't be considered.

## 7.8 GOALS AND KPIS

In conclusion, the main goals of the project are:

- **P5_G1:** Availability of relevant data for business-critical processes and data sharing across multiple regions/environments.
- **P5_G2:** Sufficient data description, relevant for regulation purposes.
- **P5_G3:** Necessity for governed and shared data across domains/geographies with tailored features for moving or querying data.
- **P5_G4:** Data access controls for moveíng or querying data
- **P5_G5:** Enforcement of data policies, for example, by masking or reduction of data
- **P5_G6:** Intelligent data movement and protection, with optimizations based on costs and business policies

Finally, important KPIs that can be monitored along the project are given by the following Table 12.

| KPI code | Category | Related Goals | Description |
|---|---|---|---|



Funded by the European Union

| | | | |
|---|---|---|---|
| P5-KPI1 | Accuracy | P5_G2 | As we want to make this metadata based, this needs to be extremely accurate. (G2-Data Governance) |
| P5-KPI2 | Accuracy | P5_G1 & P5_G5 | Data access policies for the correct parties is crucial. (G1-Tech Governance & G5-Data Governance) |
| P5-KPI3 | Time/costs saving | P5_G6 | As we have huge volumes, optimising is key in cost reduction. (G6-Tech Governance) |
| P5-KPI4 | Confidential data management | P5_G3 & P5_G4 & P5_G5 | Access compliant to the confidentiality ratings needs to be enforced. (G3-Tech Governance & G4&5-Data Governance) |
| P5-KPI5 | Data harmonisation | P5_G1 & P5_G6 | Having the right data being available for consumption and combining will intrinsically promote usage. (G1&6-Tech Governance) |

*Table 12: KPIs FOR THE SHARED FINANCIAL DATA GOVERNANCE USE CASE*



Funded by the European Union

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

## 8.1 PILOT OVERVIEW

The pilot's objective is to link sensor data from the private companies' deployment of environment and energy consumption monitoring with building energy profiles administered by public authorities. The two partners involved in this pilot are BOX2M[23], a private company, and RT[24], a public authority of Tuscany Region, Italy. The aim is to enable the reconstruction of static and dynamic energy files for public and private buildings, as well as the mapping of territorial energy efficiency and air quality trends. Open data on weather and air quality are also part of the analysis. 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.

## 8.2 DATA DESCRIPTION

The existing data, its characteristics, volume, and accessibility was described in detail in D2.1, under the section 3 of the "USE CASE PILOT #6: REGIONAL PLANNING FOR ENVIRONMENTAL SUSTAINABILITY" chapter. The synthetic data comprises the multiple datasets identified for the pilot, aggregating information regarding hydrometric, rainfall, or thermometry records, daily temperature values, air quality measurements, thermal and energy performance certificates, and environmental sensor and energy consumption data. The sizes, accessibility, and the update frequencies vary between datasets.

## 8.3 SYNTHETIC DATA GENERATION

The pilot's data is a collection of heterogeneous datasets in both format and accessibility. A share of the datasets consists of open data, publicly available through online repositories. For such data, there is no need for synthetic data generation processes, as the web repositories serve both historical and daily record updates, delivered in widely used and parseable data formats. Therefore, the strategy consists of crawling and storing the data for the project's future use. The other share of the datasets consists of private structured data that follows specific schemas. The schemas are provided by the pilot and synthetic data generation processes are needed in order to synthesise an amount of records that approximates to the real data volume handled by the pilot at a time frequency that is also specified. The fidelity of the generated data points and the correlation between them is of no measurable importance, following only that it obeys the schema and other interoperability considerations.

For all the datasets, Apache Airflow was the tool used to define, schedule, and maintain the workflows that download, store, or generate the synthetic data. The ARPAT[25] dataset gets updated on a daily basis, by querying the last recorded temperature observation from the web repository and storing it in the local filesystem, generating a uniquely timestamped identifier for the files to be saved. The same happens to the SIR_TEMP[26] dataset. For both, a GET request is sent to the respective web repository, fetching and storing the response as a

[23] https://www.box2m.com/
[24] https://www.regione.toscana.it/
[25] https://www.arpat.toscana.it/temi-ambientali/aria/qualita-aria/bollettini
[26] https://www.sir.toscana.it/archivio/dati.php?IDST=termo_max&D=json&IDS=TOS11000515



Funded by the European Union

JSON file. The RT_APE and RT_CIT datasets are synthesised based on their schemas, starting by matching the original data volume reported by the pilot, and proceeding with a daily update that mimics the approximate rate of real updates of these datasets[27]. BOX2M provides their own synthetic data, which consists of daily JSON records of environmental sensors and energy consumption data. The BOX2M dataset is updated every 15 minutes, via a call to the pilot's client API, returning a fresh record that is saved in the filesystem. BOX2M also provides their own hardware and has created the data ingestion pipeline using Apache Airflow[28].

[27] https://gitlab.teadal.ubiwhere.com/teadal-tech/airflow-rt
[28] https://gitlab.teadal.ubiwhere.com/teadal-tech/airflow.box2m



Funded by the European Union

# 9 GENERAL ARCHITECTURE

## 9.1 INTRODUCTION

In this chapter, the initial architecture for TEADAL is presented. The aim is to build a solid foundation onto which the upcoming TEADAL tools will be developed and which will be the basis for the other technical WPs to describe their specific components. In that regard, it is worth mentioning that next versions of TEADAL architecture might diverge in some aspects from this initial definition as the project evolves and the TEADAL tools get better defined.

TEADAL's architecture is presented as an interlocking of 4 views, as explained in [1]. First, we present the requirements view, which describes the architecture from the perspective of the requirements or features that deeply affect its composition. Second, the conceptual view describes an overall picture from the data-cycle perspective of the architectural components. Third, the process view depicts the internal processes that the architecture expects. And finally, the TEADAL node which presents the set of tools and components that represent current TEADAL's architecture together with the CI/CD pipeline that automates its usage. We end this chapter by sketching how our architecture fits the purpose of the evidence-based medicine use case. Following deliverables will show in more detail how each use case fits with TEADAL's architecture.

Before providing the description of the architecture, some preliminary concepts need to be explained to give the appropriate context to the reader.

TEADAL focuses on data sharing between different organisations. These organisations belong to the same federation, which provides a common set of rules for data sharing, enabling the development of federated data spaces. Thus, TEADAL provides an architecture where a set of tools enable organisations to leverage these federated data spaces for data sharing. Figure 6 shows the federated data governance, where TEADAL tools are required for data sharing among organisations.

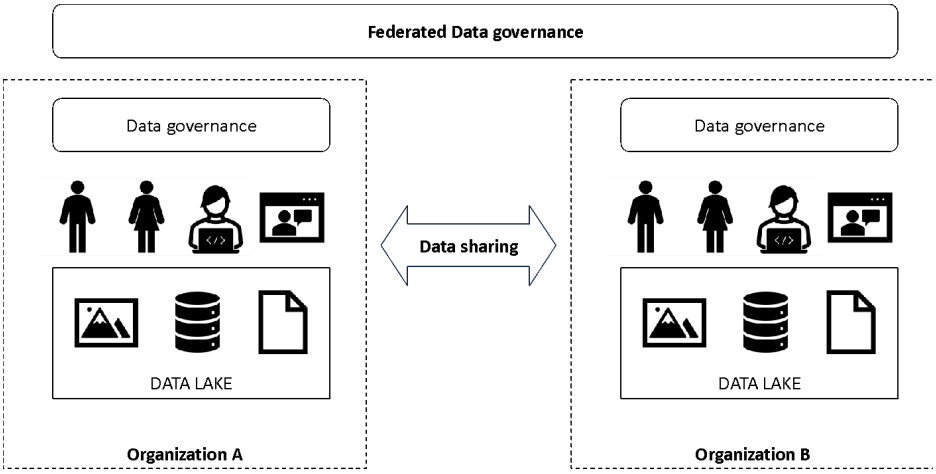


*FIGURE 6: DATA SHARING BETWEEN ORGANISATIONS, TEADAL'S SCOPE RESIDES WITHIN THE FEDERATED DATA GOVERNANCE.*

Data Mesh [2] is a design principle adopted in TEADAL's architecture. It calls for a decentralised approach to data governance and tackles the data life-cycle in large-scale organisations, bringing to the picture several key concepts for data sharing. First, the minimal unit of shareable data is the data product, which is defined by domain experts. The second concept is that the domain has the data ownership, in contrast with other paradigms, where



Funded by the European Union

data belongs to data experts of the given company, which usually lack the specific knowledge concerning data handling. Thus, domain ownership ensures maximal utility of the data. The third concept is that the data life cycle is managed through a self-service platform, implying that the domain experts will choose the tools they require from the platform to autonomously manage their data products. The fourth concept is the federated computational governance, implying that, within a federation, data products are constrained to a set of rules that enable its governance, requiring an automated enactment of policies.

TEADAL goes beyond Data Mesh, by considering its principles in the interaction between different organisations. This shifts the concept of data product to federated data product (FDP), as the minimal unit of shareable data between organisations. Furthermore, this shift triggers many of the challenges that TEADAL tackles.

Together with Data Mesh, TEADAL also embraces the Service Mesh paradigm. This builds a layer of proxies connected to TEADAL services, which intercepts service communications to enable TEADAL features, such as security, policy enforcement, or traceability.

## 9.2 REQUIREMENTS VIEW

The requirements view introduces the features that deeply affect the architecture composition. These features are extracted from non-functional requirements, which were obtained during the requirements elicitation process for each use-case, developed during the first project iteration (see deliverable D2.1). In addition, general non-functional requirements of the project are considered.

The following list introduces general non-functional requirements considered for the design of TEADAL's architecture.

1. Automate, as much as possible, data sharing between different organisations.
    a. Dynamically prepared shareable data.
    b. Ease infrastructure management.
    c. Simplify policy definition.
    d. Data discovery capability.
2. Optimise data-sharing process.
    a. Control inter-organization data-sharing overhead (Friction).
    b. Optimise data and computation placement along the stretched data lake (Gravity).
    c. Minimise the energy consumption for the data-sharing process.
3. Trust.
    a. Policy enforcement and verification.
    b. Confidentiality and privacy enforcement and verification.
    c. Ensure and verify data integrity and provenance.

Each feature is translated to a specific characteristic of the developed architecture.



Funded by the European Union

**1.a Dynamically prepared shareable data** has several implications over TEADAL's architecture. First, shareable data is not any data, but a Federated Data Product (FDP) that has to obey certain rules according to a common framework given by the federated data governance space. In that regard, an FDP is a pointer to the data, which can be accessed through the FDP REST API, also containing the policies to comply with the access to data, as well as computation capabilities to address them. Further, in TEADAL's view, shareable data is considered not only from the owner's perspective, but also from the consumer's relation with the owner. Hence, to bring this higher level of flexibility, from each type of sharing agreement between a data owner and a data consumer, the FDP has to be instantiated as a Shared Federated Data Product (SFDP), to encapsulate the specificities of the agreement. Therefore, TEADAL architecture allows sharing the SFDP between a data owner and a data consumer, and this SFDP has to be dynamically built after and according to an agreement.

**1.b Ease data lake management**. TEADAL's architecture is service-oriented, hence, components are related to services and follow the common principles of these architectures [3]. In addition, TEADAL architecture aims at hiding, as much as possible, the complexity of the data lake management to its users. Hence, when possible it introduces serverless capacities so that the users do not need to care about the resource provisioning or other data lake operations, as the management and orchestration of resources will be dealt with by TEADAL's control plane.

**1.c Simplify policy definition**. Each FDP might require a different set of policies to allow its sharing. Further, these policies need to be made understandable and verifiable by the data owners who define them, while also being readable and enforceable for the automated policy enforcement points. Consequently, TEADAL's architecture provides the required software components to enable these policies' translation from human understandable language to machine readable code, at the enforcement points.

**1.d Data discovery capability.** TEADAL sharing space is common amongst different organisations, therefore their members need to be able to discover the available FDPs, without accessing the data, to preserve privacy and confidentiality. To that end, TEADAL architecture incorporates TEADAL Data Catalogue to provide this visibility and discoverability to all federation members.

**2.a Control inter-organisation data-sharing overhead (Friction)**. TEADAL's advancement over data sharing between different organisations brings additional requirements to FDPs, in order to be shared. This implies that FDPs will have specific policies that affect how data is shared, leading to additional data processing steps. TEADAL architecture has to manage these steps transparently to the user, while being aware of these extra effort to control and minimise it whenever possible. TEADAL has defined friction as a set of data pipeline steps (see D3.1), which will be measured and controlled by TEADAL's control plane. Additionally, the overarching requirement of optimising the data sharing process requires TEADAL to incorporate performance monitoring tools.

**2.b Optimise data and computation placement along the stretched data lake (Gravity).** TEADAL assumes that data lakes can be stretched along the computing continuum. This defines a new dimension where data and computations can be placed, given that, depending on the system configuration, the optimal placement for data and computations can vary. In specific situations, it is more convenient to process data next to its source. However, devices next to data sources are usually constrained, therefore, a distribution of the computational tasks is required. TEADAL's architecture considers these cases and leverages TEADAL's control plane to manage this type of trade-offs. See deliverable D4.1 for further details on the stretched data lake and the control plane.



Funded by the European Union

**2.c Minimise the energy consumption for the data-sharing process.** TEADAL's objective of minimising its energy footprint is holistic. This is also considered from an architectural perspective. In that regard, TEADAL's architecture uses serverless technology whenever is reasonable, to benefit from its scale-to-zero capacity [4], which means that whenever services are not needed, there is no need to have infrastructure provisioned and running.

**3.a Policy enforcement and verification.** TEADAL enables data sharing among organisations that belong to the same federation. Although this pre-assumes a certain level of trust between them, TEADAL has to provide the tools to ensure that all processes are defined as agreed. TEADAL leverages Service Mesh proxies attached to the FDPs that can intercept any request to verify that the policies are being fulfilled, building a network of interception proxies. TEADAL trust plane implements blockchain technology to keep track of all interactions, enabling verification and control processes. See deliverable D5.1 for more details on the trust plane and related tools.

**3.b Confidentiality and privacy enforcement and verification.** TEADAL architecture addresses privacy and confidentiality from two perspectives. On the one hand, the use of an FDP (or an SFDP) as the data sharing entity allows the data owner to define the data visibility for the data consumer so that the latter can only access data from an interface (REST API) defined by the data owner, which is enforced thanks to the Service Mesh capabilities. On the other hand, TEADAL architecture builds pipelines from the dataset to the FDP, and then, to the SFDP that enforce confidentiality and privacy. Further, prospective TEADAL tools such as privacy preserving computations will ensure privacy along these pipelines. Finally, the trust plane will also monitor these pipelines to track all processes and provide the data owner with verifications tools.

**3.c Ensure and verify data integrity and provenance.** TEADAL's architecture, specifically through its control plane and blockchain technology, provides traceability of data access and manipulations through the data life cycle. It captures and stores evidence, which allows federation members to audit the system.

## 9.3 CONCEPTUAL VIEW

The high-level conceptual view of the Teadal architecture comprises multiple components and entities that are used throughout the technical work packages. The core parts and their interplay are explained in the following sections: the responsibilities of each component, the interfaces through which they communicate, and how they interact to fulfil their requirements.

In its simplest form, data exchange consists of two entities: a data provider and a data consumer. Within Teadal, we only consider cases where they belong to different organisations, otherwise data could be exchanged with the help of a regular data product, i.e., one that is internal to one organisation. In all cases, Teadal can be the tool that supports the data exchange in a controlled way: (1) it identifies and resolves data gravity and frictions that hamper the exchange process; (2) it assures policies that govern the exchange; (3) it provides all the resources required for processing and storing the data; and (4) it gives evidence of how data is transformed. The third point, in particular, leverages the properties of serverless computing to enable trustworthy data sharing with minimal operational overhead.

To provide these capabilities to the provider and the consumer, we present the following life cycle for a FDP. It consists of five phases that are centred around the data product, i.e., from its provision until its removal:

1. **Data Onboarding**: When a provider shares its data products with other federation members, the Teadal platform is in charge of transforming them into one cohesive



Funded by the European Union

logical product, called FDP, irrespective of their physical distribution within the federation. Data products can be enriched with policies (1a), which are supplied by domain experts. During the onboarding phase, the majority of the policies revolve around storage considerations, restricting its physical location within the federation. Other types of policies can include serverless functions (e.g., transformations or access restrictions). These functions are shared in a repository (1b) and any following policy specification must only include a reference to them.

2. **Publishing**: The data provider makes the federated data product accessible to the data consumer by registering it through a federation-wide catalogue (**2a**). Although the catalogue operates as a single unified entity, effectively, it can be decentralised within the federation, for example, through a distributed database. The catalogue contains the accompanying metadata for each FDP, including the assigned policies. By accessing the catalogue, the data consumer can discover data products that match their specific requirements, e.g., in terms of quality or quantity.

3. **Sharing**: As soon as the data consumer has decided on an FDP (3a), all parties involved (i.e., provider and consumer) must reach an agreement on how data will be shared (3b). This agreement involves all policies assigned to the data product; nevertheless, it can also be further extended, e.g., with custom data transformations or restrictions detailing where in the federation the transformation must occur. These policies are again formulated as functions that are distributed within the platform. Thus, in this phase, the data processing pipelines are prepared, which comprise all transformations of federated data products between the provider and consumer. The result is an SFDP, to be consumed in the following phase. Agreements themselves are stored by all parties involved and serve as proof of trust among partners.

4. **Consumption**: With the agreement as proof, the data consumer can request access to the federated data product (4a) through the federation catalogue. Requests are validated (4b) and forwarded to the FDP (4c) by the data provider's query interface. The control plane now optimises the execution of serverless functions by distributing them over the computing infrastructure, if not specified otherwise in the agreement. Data is then transformed through the installed pipelines (4d), according to the assigned policies. Due to filtering mechanisms, transformations, and other functions executed on the FDP, the data received by the consumer (4e) is most likely not identical to its initial copy. In fact, the consumer's local instance (or rather, its view on the data product), is called the SFDP. Current TEADAL's vision builds one SFDP for each FDP, hence the consumer accesses as many SFDPs as FDPs they have obtained access. After transforming the data, all processing resources that were used can be freed, i.e., due to the serverless computing's capacity to scale to zero.

5. **Discontinue**: The agreement can be concluded due to different reasons, e.g., when the maximum number of accesses or a time limit has been reached or if one of the parties withdraws from the agreement. Under any of these conditions, the agreement ceases, and the respective federated data product cannot be accessed anymore. Consequently, the control plane releases all resources associated with consumption, i.e., processing resources for running serverless functions, but also any consolidated storage that improves the data consumption. To that extent, the control plane makes use of the data lineage capabilities provided by the trust plane to identify all components that need to be terminated.

This concludes the lifecycle of a federated data product. In the following, we will describe core components of the presented architecture, this includes first and foremost, the FDPs and the catalogue, as well as the federation's underlying control plane, trust plane, and the



Funded by the European Union

security service mesh. For further details on these entities, see deliverables D3.1, D4.1, and D5.1.

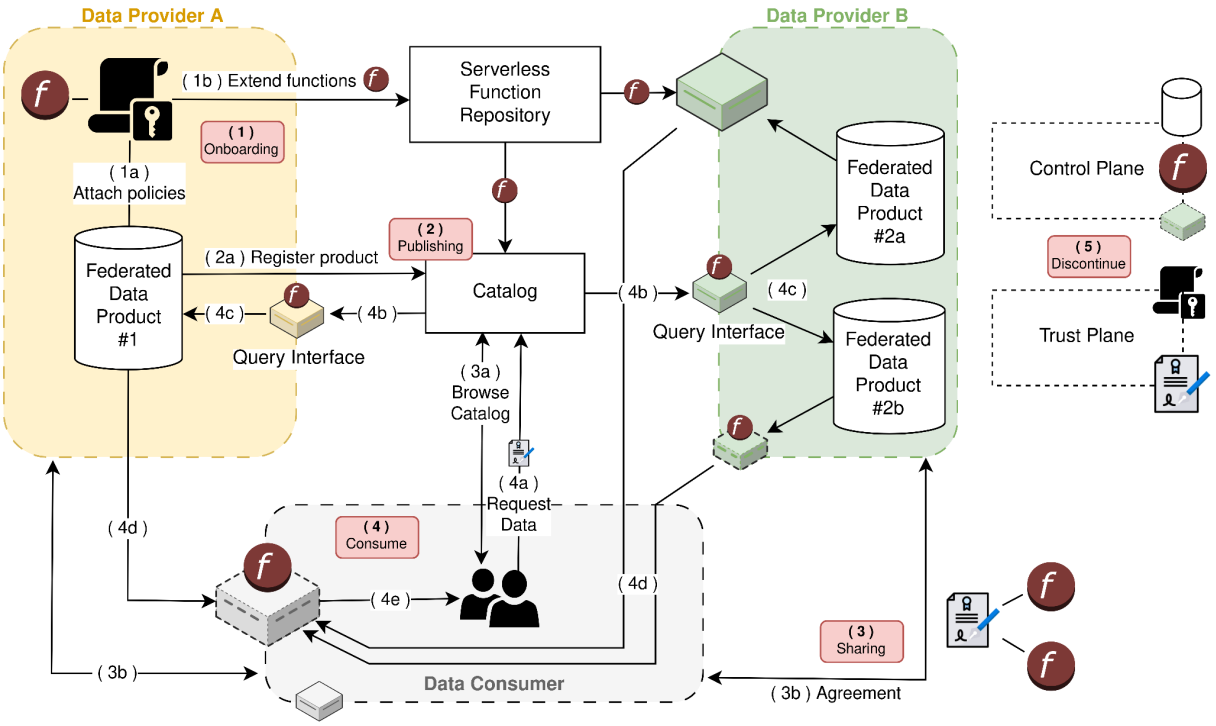


*FIGURE 7: CONCEPTUAL (COMPONENTS AND PROCESSES) VIEW OF TEADAL'S ARCHITECTURE.*

**Federated Data Product**: Consider a data product provided by the federation members, where domain experts are in charge of describing the product with usage policies. This concept is already known from data meshes. The data can be stored on the premises of the data provider, or on resources provided by the federation, this is the responsibility of the control plane.

Bundled with its assigned policies, the data product can be registered in the Data Catalogue and it is now federated within the Teadal platform. The FDP is then exposed for consumers in the Teadal federation, which can access it through the provided interfaces (REST API). For every consumer (group), the FDP is going through a custom transformation process (also called FDP-SFDP-pipeline). The distinct transformation steps are determined by the usage agreement between the data provider and consumer. The instance of the FDP which is accessible by the consumer is called the shared federated data product (SFDP). This is illustrated in Figure 8 and deliverable D3.1 covers this aspect in more detail.

**Federation-wide Catalogue**: Must be accessible by all federation members. However, federated data products can have custom visibility settings that hide them from unauthorised groups. The catalogue is the gateway for the data consumers and providers, which masks various platform details, e.g., the precise physical location of FDPs. By accessing the catalogue, potential consumers can search for data sets that fulfil their requirements, e.g., in terms of metadata, quality, or quantity, enabling the discovery capacity within the federation.

**Trust plane**: The trust plane is a critical component of a federated data governance architecture, playing a significant role in ensuring security, privacy, data integration, and data governance. It helps define and enforce access control policies, authentication, and authorization mechanisms to ensure that only authorised users and services can access data. It can also track changes and usage patterns, which is crucial for troubleshooting. The trust plane in Figure 7 can work in four parts of the most relevant interactions. Before a



Funded by the European Union

provider shares its data products with other federation members, the trust plane ensures that providers can access them. By verifying the legal aspects of sharing, the trust plane protects the FDP from unauthorised and noncompliant behaviours. This validation process is a critical step in maintaining the security and integrity of the data product within a federation. Once the initial check is validated, a set of access policies (legal constructs, such as audit procedures or penalties for misbehaviour) are embedded while sharing metadata (1a) or serverless functions (1b). However, the provider is responsible for cross-checking whether all access policies are properly embedded during publishing or registration (2a). Whenever a consumer requests access to a catalogue (4a), the trust plane verifies it, and will only allow browsing access (3a) if all the access policies established by the provider have been met. Providers need to match consumers' requirements and provide policy-specific agreements along with access to data products. The trust plane ensures that these agreements are compliant during FDP consumption (4d). It will be alerted immediately to the control plane when it identifies misbehaviors or infractions by either the provider or consumer. More about the trust plane can be found in D5.1.

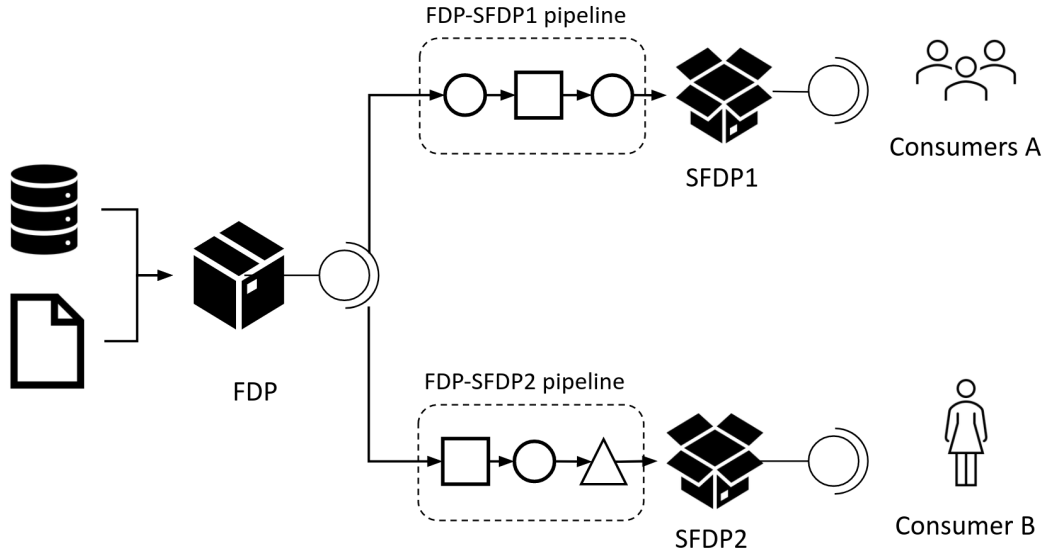


*FIGURE 8: FDP TO SFDP PIPELINES. EACH SFDP IS BOUND TO A SPECIFIC AGREEMENT BETWEEN THE DATA OWNER AND THE CONSUMER.*

**Control Plane**: The control plane is responsible for providing management over the FDPs and the physical resources of the stretched data lakes. It hides the complexity of running components of the data lake in multiple locations, and uses the Kubernetes control plane API to orchestrate workloads and resources. These orchestration decisions are obtained by mapping workload requirements with computing or storage capacities. In the next iterations, the control plane will use results from other TEADAL tools, e.g., tools that consider the energy consumption, to optimise the overall performance of the data lake.

**Service mesh & security**: Augmenting system functionality through message interception is a key tenet of a service-oriented mesh architecture. Both inbound and outbound service communication transit, through a network of proxies, isolate services from each other and the rest of the network. On intercepting a service request, a proxy can inspect it, decide whether to route it to the target service and possibly alter it before routing it. Likewise, proxies intercept service responses and possibly process them before forwarding them to service clients. The mesh can leverage this interception mechanism to enrich service functionality without requiring any alteration to service code. In particular, TEADAL service mesh helps with the security of data assets, it enables tracking the FDP and SFDP life cycles to produce verifiable evidence, and it helps with the observability of complex metrics such as gravity and friction.



Funded by the European Union

TEADAL employs a mesh infrastructure to provide data product access control. Data product services (FDP, SFDP) do not need to implement access control. The mesh provides it by intercepting requests to data product services and delegating access control to the following components, which are implemented and deployed independently of data products:

- *Policy decision point*. Given a service request, it decides whether to allow it. The decision process entails evaluating access control policies applicable to the service request. Policies are written in a high-level domain-specific language.
- *Policy store*. It allows product owners (or someone on their behalf) to store and manage the policies for their respective data products as well as making them available to the policy decision point for evaluation.
- *Policy enforcement point*. It interacts with the policy decision point to determine whether to allow or deny service requests and with the mesh proxy to enforce the access control decision.

## 9.4 PROCESS VIEW

This section provides an overview about the processes needed by the TEADAL architecture. The processes are represented as they are currently envisioned, but they might be subject to change in future, as they are still being defined. In addition, the section presents the actors that have a role in the TEADAL's architecture, it provides more details about the network interception mesh and the policy enforcement process.

### 9.4.1 TEADAL ACTORS

The first iteration of the TEADAL architecture considers five different actors or roles that can intervene in the processes. The following is a short introduction on them. For more details, see Deliverable 5.1.

Data Lake Operator (DLO) refers to the actor able to install the TEADAL tools, and the one managing security and compliance of the installed environment.

FDP Designer (Designer) refers to the actor defining policies, specifications, and metadata required for creating an FDP.

FDP Developer (Developer) refers to the actor that implements the software components to develop an FDP. The FDP Developer is also responsible for ensuring compliance with TEADAL rules.

FDP Provider (Provider) refers to any actor who represents an organisation of the federation with the right to access and share data. The three previous actors can also be considered Providers.

FDP Consumer (Consumer) refers to any actor who represents an organisation of the federation that searches for an FDP and negotiates the agreement to access the SFDP. Finally, the Consumer is responsible for developing client-side code needed to interact with FDPs or SFDPs.

### 9.4.2 TEADAL REGISTRATION

Each member of a TEADAL-powered federation has to follow a sequence of steps, as defined in Figure 9, to properly register itself within the federation.

The DLO interested in being part of a TEADAL's-powered data-sharing federation has to install TEADAL tools in its data lake, which are packed in the TEADAL node. The correct installation of TEADAL tools will register a new federation node in the TEADAL's control plane.



Funded by the European Union

Then, to join the federation, the DLO will have to sign the federation agreement, which contains a common set of rules for data sharing within the federation.

Once the agreement is signed, the DLO will register its data lake using the tools provided by TEADAL. The trust plane is in charge of verifying the signed federation agreement.

Once this process is completed, the DLO can start sharing its data from its data lake with all federation members.

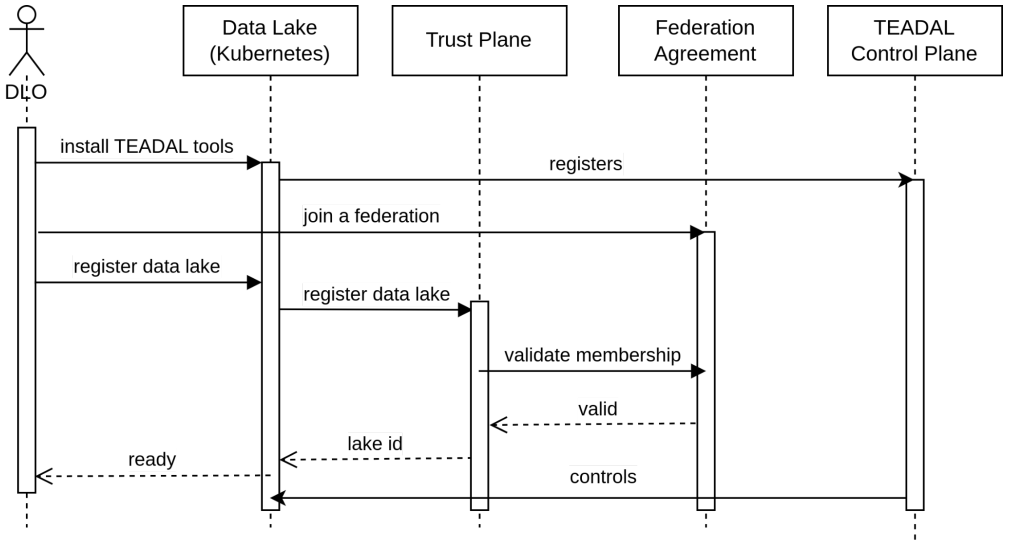


*FIGURE 9: TEADAL REGISTRATION PROCESS.*

### 9.4.3 TEADAL FEDERATED DATA PRODUCT

TEADAL's key enabler for federated data sharing is the FDP. As explained above, this will have a REST API to expose data, a description of the policies, and a computation recipe.

Figure 10 shows the process to develop and make an FDP available for the TEADAL-powered data-sharing federation. First, we assume that the DLO might not have the technical skills to design the FDP. However, the DLO needs to provide clear instructions in terms of the policies that the FDP requires and the capabilities that the interface needs according to their envisioned use case for the FDP. A designer will take the high-level needs of the DLO and develop an FDP specification for a developer to build it.

The developer will create and deploy the FDP. The FDP will reach TEADAL control plane, which will send its description and characteristics to the trust plane for validation. Furthermore, the trust plane will also start the tracking of the FDP.

Once the FDP is validated, the control plane will deploy the FDP into the federation, and the federation catalogue will make the new FDP available for browsing by the federation. As we will detail in the next phase, this does not make data available to the federation members, but allows them to know what data they can access, and if needed, request the FDP.



Funded by the European Union

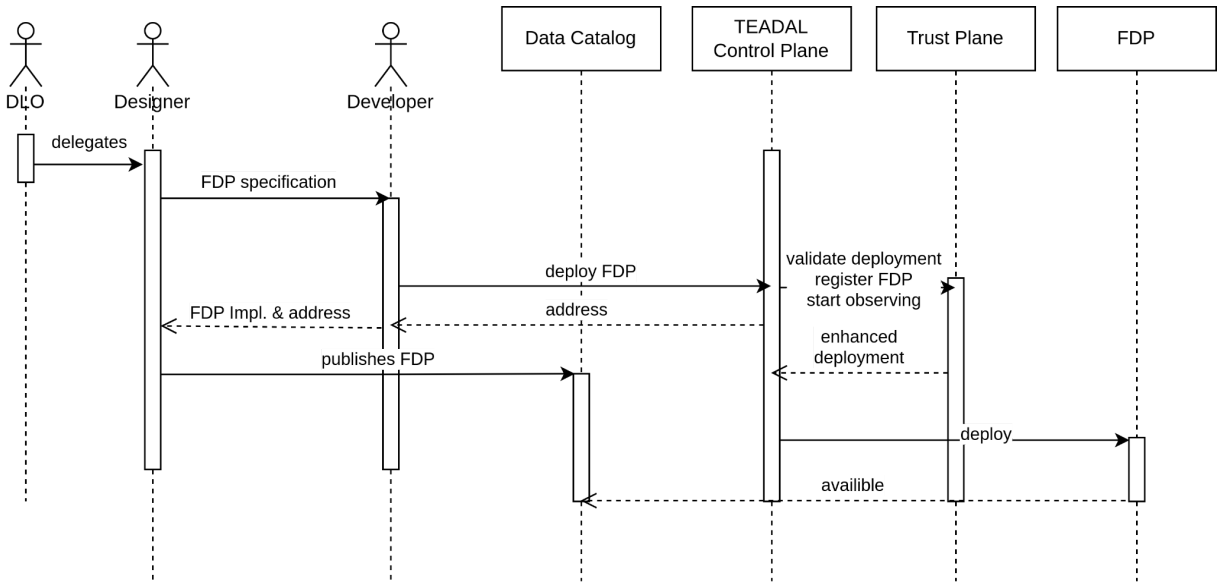


*FIGURE 10: TEADAL'S FEDERATED DATA PRODUCT DEVELOPMENT.*

### 9.4.4 TEADAL SHARED FEDERATED DATA PRODUCT

This subsection explains the process for a data consumer to gain access to an available FDP in the federation catalogue, illustrated in Figure 11.

The consumer will find its desired FDP by browsing TEADAL's catalogue, which shows metadata from the FDPs, giving consumers the understanding of what the FDP contains. The consumer will have to negotiate and sign an agreement with the FDP owner. During this process, the data owner or the consumer may introduce new requirements for accessing the FDP.

Once the agreement is signed, the Data Catalogue gives the instruction to TEADAL's control plane to deploy the SFDP, which is the instance of the FDP with specific requirements and policies of the signed agreement. TEADAL's trust plane will validate the SFDP deployment and start observing it, which will lead to the final deployment of the SFDP.

Finally, the consumer will obtain the address of the SFDP to access the data through its REST API.

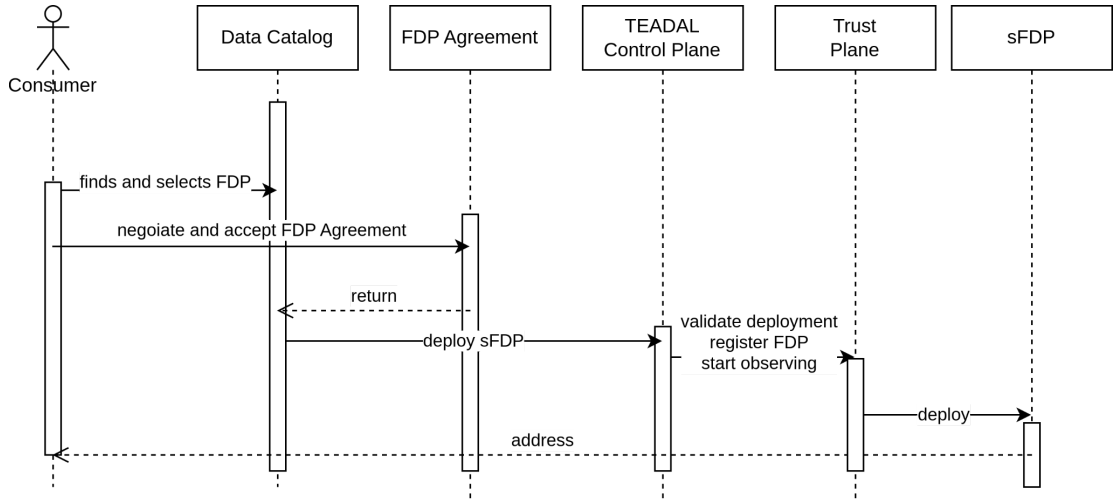


*FIGURE 11: TEADAL'S SHARED FEDERATED DATA PRODUCT DEVELOPMENT.*



Funded by the European Union

### 9.4.5 FEDERATED DATA ACCESS

At this point, the consumer has obtained an agreement with a data owner and has the address of the SFDP. Now, to access the data from the SFDP, the process follows Figure 12.

The consumer will send a data request to the SFDP, after which the SFDP will ask the consumer for authentication. To comply, the consumer will login through the access control component and, at that moment, TEADAL's trust plane will get the consumer's reference.

The login will grant a token (specifically a JSON Web Token - jwt) to the consumer. The consumer will proceed to request the data again, but now in possession of the token. The access request is sent to the trust plane, which will forward the request for data to the FDP. At this point, all required policies will be applied to the data at the FDP, the FDP will issue a proof of the applied filters, and finally, send the data to the consumer.

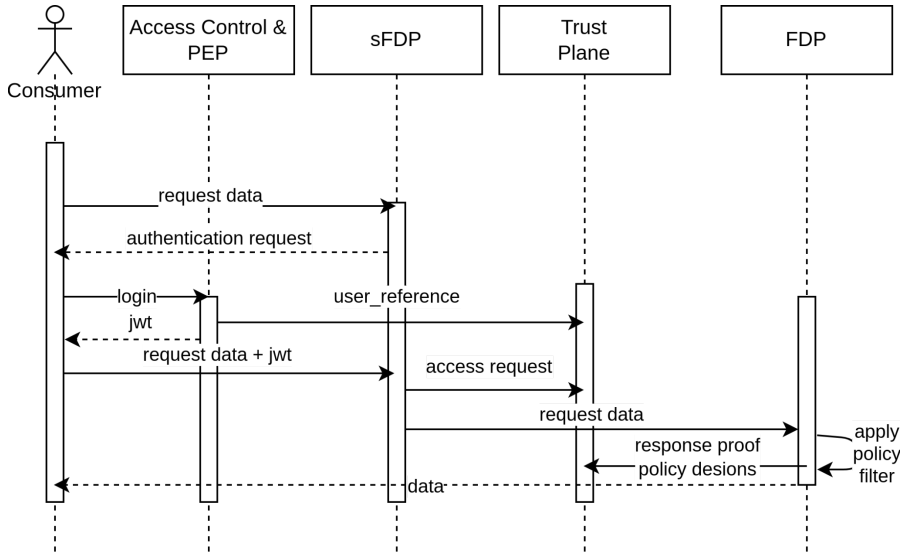


*FIGURE 12: DATA ACCESS PROCESS FOR A TEADAL'S CONSUMER.*

The process of augmenting data product functionality with access control is as follows. First off, the mesh proxy intercepts the request which the data consumer makes to the data product API. The proxy then asks the policy enforcement point to process the request. In turn, the policy enforcement point asks the policy decision point to check whether the request is allowed to proceed. The policy decision point looks up the policies applicable to the given request in the policy store and then evaluates them against the request. If the evaluation outcome indicates that the request should be allowed, the policy decision point informs the policy evaluation point accordingly. On receiving an "allow" decision, the policy enforcement point instructs the proxy to route the request to the data product service, collect the response and forward it to the data product consumer. On the other hand, in the case of a "deny" decision, the policy enforcement point issues an unauthorised error response that the proxy forwards immediately to the data product consumer without proceeding to invoke the data product API. Figure 13 illustrates the access control process just outlined in the case of an "allow" decision.

Note that both the data product and the consumer are unaware of the interception proxy. From the consumer's perspective, the request is a direct message to the data product API as if the consumer were invoking the API without a proxy in between. Similarly, the data product API processes the request as if it originated directly from the consumer and produces the same response it would if the proxy did not intercept the incoming request.



Funded by the European Union

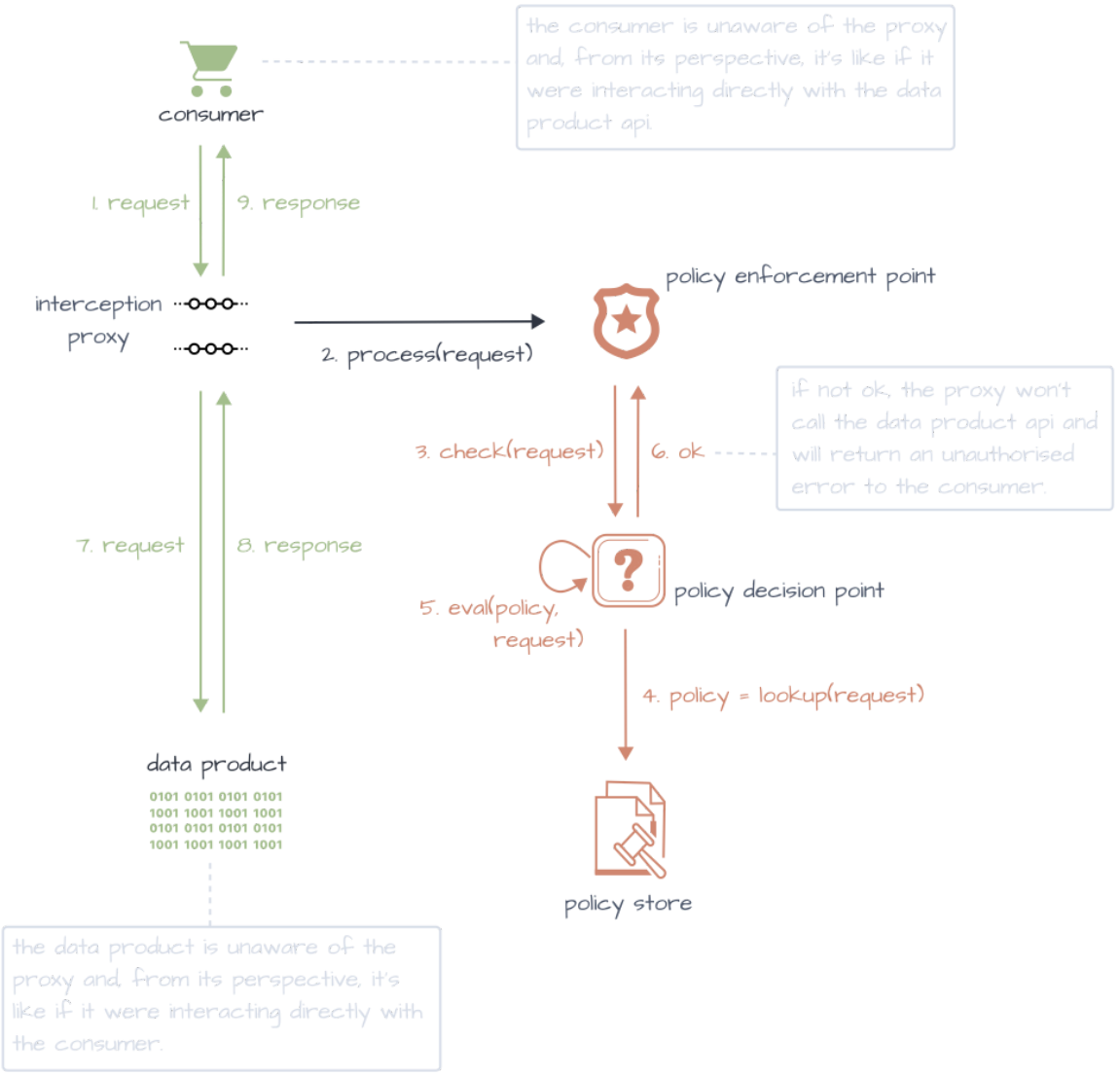


*FIGURE 13: ACCESS CONTROL THROUGH THE SERVICE MESH.*

## 9.5 TEADAL NODE

The TEADAL node is a composition of tools and components, which are exposed as one conglomerate to its users to support data exchange. The capabilities of Teadal, including those of each component, are provided through the Teadal nodes, where the core logic is combined and run. The TEADAL node includes all tools for data providers and data consumers alike. There is no distinction in the tools provided, otherwise it would require changing the platform to change the role. Most of the components used in the Teadal node have already been mentioned. The following explanation of the node will thus focus more on how the components are provided as one coherent platform. Figure 14 shows the different functionalities provided in the TEADAL node, these functionalities are aligned with the conceptual zones defined by TEADAL, see Deliverable D3.1 for further detail.



Funded by the European Union

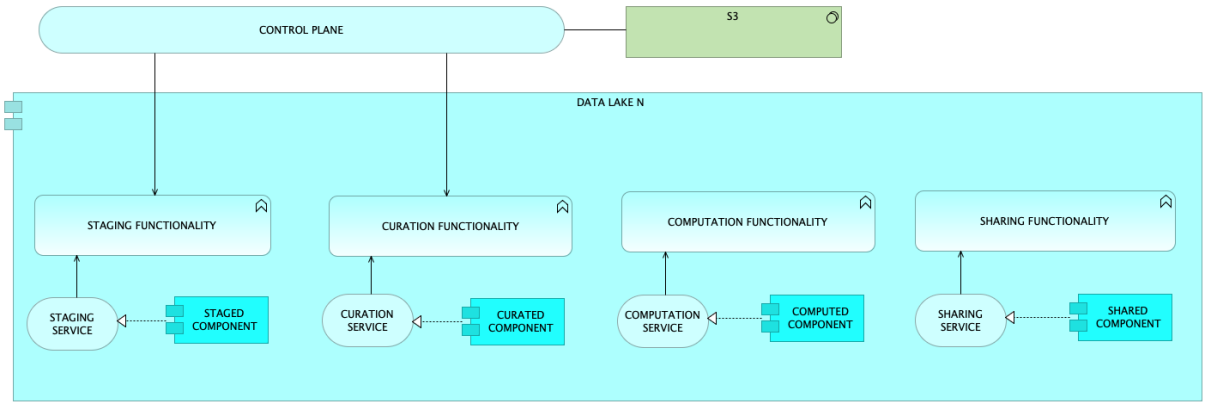


*FIGURE 14: FUNCTIONALITIES IN A TEADAL NODE.*

The tools provided by the TEADAL node can be used in arbitrary order. Nevertheless, its main purpose is to support the transition of data from an organisation (data owner) to the SFDP, through the four zones presented next. The TEADAL node contains a control plane, which is the interface to the data lake and manages data through the zones.

**Data staging zone**. This zone takes raw datasets and ingests them in the TEADAL node. To do so, this phase requires ingestion functionalities, which will move and curate the data before it is stored.

**Curated data zone**. Once ingested, data is stored permanently, at this stage data can only be accessed internally to the organisation, it is not yet part of the federation. Therefore, at this phase, storage functionalities are given.

**Computation zone**. In this zone curated data can be further analysed or modified according to the organisation needs. Consequently, this zone provides computation functionalities.

**Data sharing zone**. This zone provides all required sharing capabilities. This is the phase where the FDPs are built, made visible and available for sharing in the federation through the SFDP artefacts. The sharing zone includes all of the required functionalities of the FDP and SFDP, including computational and storage capabilities. However, due to accessibility and visibility constraints these are logically separated from previous storing and computational capabilities. Additionally, discoverability, security, and trust functionalities are key elements of this last set of sharing requirements.

### 9.5.1 Deployment

In the evolving landscape of technological advancements, the importance of detailed testing environments — referred to as "TESTBED sites" — cannot be overstated. These sites, equipped with cutting-edge resources, play a pivotal role in ensuring that developments meet rigorous standards of performance, security, and scalability. This chapter delves into the intricate details of the specific resources associated with various TESTBED sites. By cataloguing these resources, we aim to provide clarity and direction for developers, researchers, and stakeholders alike.

Figure 15 represents a CI/CD (Continuous Integration and Continuous Deployment) workflow that facilitates the development, integration, and deployment processes within a Kubernetes (K8s) cluster environment.



Funded by the European Union

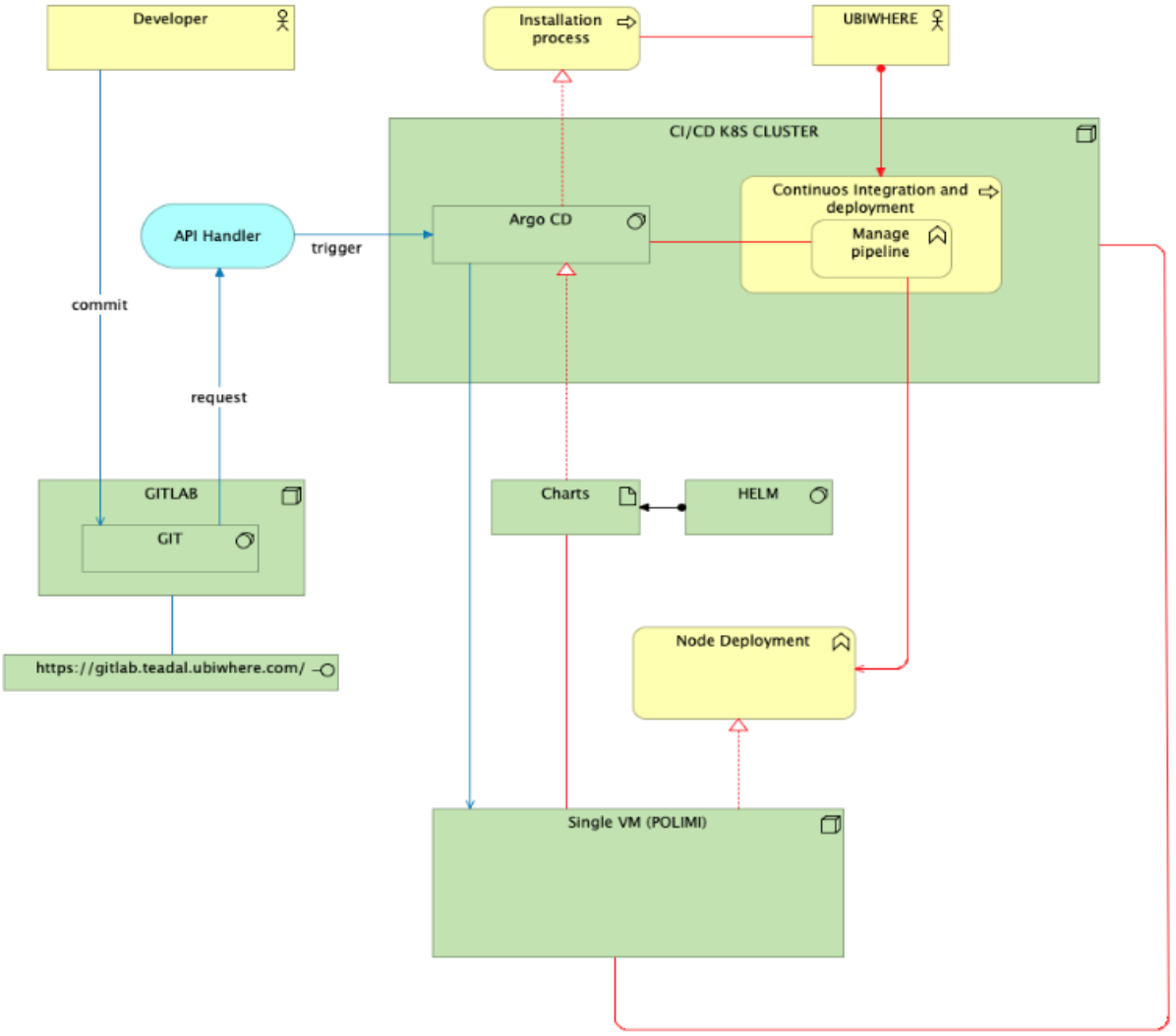


*FIGURE 15: TEADAL CI/CD WORKFLOW.*

- **Developer and UBIWHERE:**
  - The workflow begins with the 'Developer', who interacts with the platform provided by 'UBIWHERE'.
  - The developer commits code changes.
- **GITLAB & GIT:**
  - The committed code is then pushed to a GIT repository hosted on 'GITLAB'. The URL https://gitlab.teadal.ubiwhere.com/ suggests the hosted location of the GIT repository.
- **API Handler:**
  - The 'API Handler' acts as an intermediary component that receives a 'webhook' from GitLab when code is committed. This webhook is a HTTP callback that is triggered by specific events. When a commit is made in GitLab, it sends a request to the API Handler with the details of the event.
  - The API Handler subsequently triggers Argo CD operations to synchronise the application state.



Funded by the European Union

- **ArgoCD:**
    - Argo CD is a declarative, GitOps continuous delivery tool for Kubernetes. It leverages Git repositories as a source of truth for defining the desired application state. When the API Handler receives the request from GitLab, it triggers Argo CD to synchronise the application state in the Kubernetes cluster with the new state defined in the Git repository. This means that Argo CD automates the deployment and ensures that the application's state matches the state defined in the Git repository.
- **CI/CD K8s CLUSTER:**
    - This is a dedicated Kubernetes cluster where the CI/CD processes take place.
    - Inside this cluster, we see the 'Continuous Integration and Deployment' process with a 'Manage pipeline' section, indicating the various stages the code passes through.
- **HELM & Charts:**
    - HELM is a Kubernetes package manager. It uses 'Charts' for defining, installing, and upgrading even the most complex Kubernetes applications.
- **Node Deployment:**
    - Once the application is packaged by HELM and is ready to be deployed, it gets deployed on a node within the cluster.
- **Single VM (POLIMI):**
    - The diagram also indicates a 'Single VM' labelled 'POLIMI', suggesting that there's an isolated virtual machine potentially used for specific tasks or testing purposes.

The entire setup aims to streamline the development process, enabling developers to integrate and deploy their code changes efficiently using modern CI/CD practices in a Kubernetes environment. Given the detailed nature of the CI/CD processes depicted, it seems apt to include this diagram in the "TESTBED Site" Resources And Testbed Matrix" section, as it provides insights into the technical resources and workflows used for integration and deployment.

### 9.5.2 TEADAL Components

The presented ArchiMate view is depicted in Figure 16 and showcases the process and architecture involved in the creation of an SFDP. Let's break down the flow:



Funded by the European Union

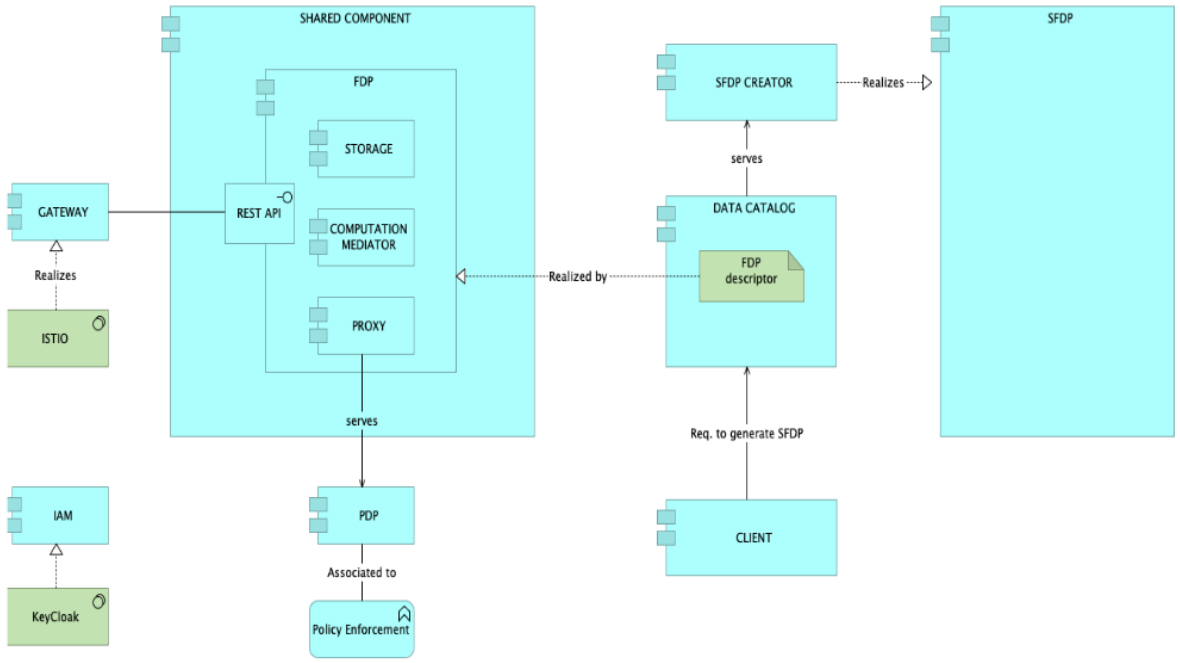


*FIGURE 16: ARCHIMATE OVERVIEW OF THE SFDP CREATION PROCESSES AND ARCHITECTURE.*

1. **Overview:**

   The diagram delineates three primary sections:

   - The Shared Component, which encapsulates core services related to data storage, computation, and interfacing.
   - The SFDP Creation Process, which involves the SFDP Creator interfacing with the Data Catalogue and the eventual realisation of the SFDP.
   - The foundational services that include the Gateway, IAM, and PDP which provide foundational capabilities such as identity and policy enforcement.

2. **Shared Component:**

   Within the Shared Component:

   - The REST API serves as the central interfacing component.
   - Storage is where all data pertinent to this process resides.
   - Computation Mediator is responsible for orchestrating the computation tasks.
   - The Proxy acts as an intermediary that serves requests, likely directing the flow of data or ensuring secure access.

3. **SFDP Creation Process:**

   The process starts with a Client who initiates a request to generate an SFDP.

   - This client interacts with the Data Catalogue to find an appropriate FDP Descriptor.
   - Post this, the SFDP Creator kicks into action. It takes the chosen FDP Descriptor, processes it, and then realises (or creates) an SFDP.

4. **Foundational Services:**

   - Gateway serves as the primary access point, realised by Istio, an open-source service mesh.



Funded by the European Union

- IAM (Identity and Access Management) is facilitated through KeyCloak, ensuring proper authentication and authorization.
- PDP (Policy Decision Point) is intricately linked with the Policy Enforcement component, ensuring that data shared in the federated environment adheres to agreed-upon policies.

In essence, this view captures the journey from a client's desire to create an SFDP, through the retrieval of a descriptor from the Data Catalog, to the actual creation of the SFDP, all while ensuring that foundational services guarantee security, policy adherence, and efficient data flow.

### 9.5.3 TEADAL Tools

The last facet of the TEADAL node section is the detailed list of tools that are included in the TEADAL node as a baseline for future contributions.

| Name | Category | Functionality |
|---|---|---|
| ArgoCD | DevOps | GitOps IaC tool for Kubernetes clusters. |
| Istio | Networking | Service mesh network. |
| Keycloak | Security | Access/authentication manager |
| OPA | Security | Policy enforcement tool |
| Reloader | Security | Check changes in config files |
| MinIO | Storage | Object storage |
| PostgreSQL | Storage | SQL database |
| Grafana | Monitoring | Platform for data visualisation |
| Prometheus | Monitoring | Tool for collection and storage of computing metrics |
| Kiali | Monitoring | Istio console to monitor and control the service mesh |
| Jaeger | Monitoring | Tracing tool to map data flows and requests |
| httpbin | Apps | Service to test http requests, for testing |
| Airflow | DAG | Management platform to define pipelines and workflows on data |
| Kubeflow | DAG | Management platform to define pipelines and workflows on data related to machine learning |
| Kubestellar | Control plane | Manage running TEADAL workloads across the locations/clusters in TEADAL |

*TABLE 13: LIST OF TOOLS FEATURING IN THE TEADAL NODE BASELINE.*

Finally, the TEADAL node also includes a "dummy-FDP" which is an example of a federated data product.



Funded by the European Union

## 9.6 ARCHITECTURE FITNESS FOR PURPOSE

The goal of TEADAL is to allow different organisations within the same federation to share data efficiently while ensuring trust and data privacy. By extending the Data Mesh concept and leveraging a Service Mesh, TEADAL architecture sets the required baseline to build the needed TEADAL tools to accomplish its objectives.

The TEADAL architecture has to fit many applications such as healthcare, environmental sustainability, industry 4.0, mobility applications, etc. Here, we provide the healthcare scenario, where dataset complexity is highly diverse, as a small sample to show how the architecture fits this use case. Please, consider that this is an initial draft about how the architecture can fit one of the TEADAL's use cases, so misalignments could be found. In following deliverables, this section will be more detailed and include other use cases from the TEADAL consortium. Workshops are being conducted with each use case partner to validate the technical ideas and to integrate them in the specificities of the use case.

Generally, hospitals require a lot of effort to select and prepare the data in accordance with internal regulations and general norms (e.g., GDPR), as well as common data formats (e.g., OMOP), and mutual agreement on semantics (e.g., SNOMED). The high heterogeneity of patients data and the regulatory constraints create difficulties in finding and combining large pools of diverse patient data in a timely manner. Our architecture is capable of providing all the required components to satisfy the use case proposed in Figure 17.

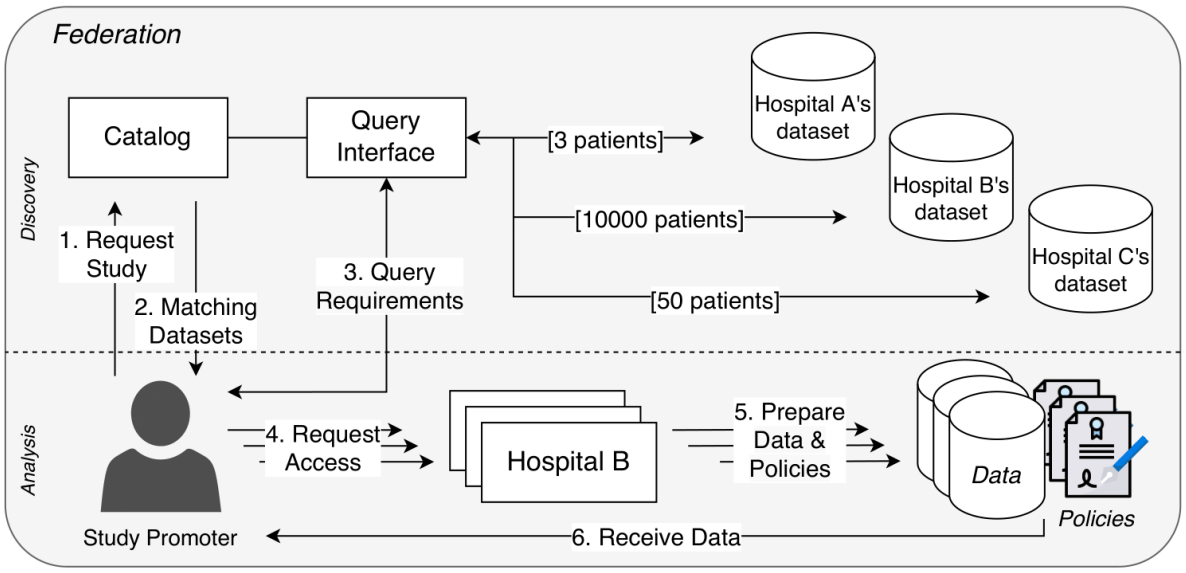


*FIGURE 17: STUDY PROMOTER WORKFLOW.*

***Searching for relevant patient data***: In this phase, the study promoter first needs to request the data specified in the Data Catalogue. By using metadata (e.g., data types, usage consent) in the Data Catalogue (established in the data mesh), study promoters can search through FDPs. According to the search query requirements, several FDPs can provide the needed data. The study promoter uses this SFDP for further use in a subsequent analysis, along with the agreement made with the data providers. Data access rules are negotiated between the data provider and the study promoter, during this agreement.

***Analysing actual data***: Upon reaching agreement with the FDPs owners (which might include different organisations), the promoter can request the actual data, accessed through the SFDP. Providers need to ensure that only relevant data is available to study promoters.



Funded by the European Union

Following the agreement, all formats must be converted according to legal obligations and federation guidelines. TEADAL control plane will prepare the required pipelines achieving the expected resource usage sustainability in the data lake, by carefully selecting the location of computations. Furthermore, it provides support to the TEADAL trust plane for the access control, according to the federation and the agreement rules. TEADAL trust plane ensures compliance with privacy and confidentiality of the data.



Funded by the European Union

# 10 CONCLUSION AND FUTURE WORK

This deliverable gave an updated overview of all of the pilot use-cases of the TEADAL project, with a focus on the synthetic data generation and the changes in requirements identified from the deliverable D2.1. Due to the detailed nature of D2.1, the changes in most pilot use-cases were fairly incremental, even as the pilots are getting more fleshed out and solidified. A more detailed final description of the pilot requirements and use-cases will be given in deliverable D2.3, which will provide analytics of the pilots, and necessary changes and additions identified after the first test deployment iteration.

From the pilot use-cases side, an important contribution of this deliverable is the detailed description of the shared financial data governance pilot, which was missing from D2.1. The pilot case is still missing a few details for data synthesis. This is currently being worked on by the pilot partner.

Data generation for the TEADAL project has proved to be a more varied process than originally intended, with different parties taking the responsibility of providing the data, different privacy and confidentiality levels for different datasets, different sources, and different needs for data integrity. Table 14 gives an overview of what kind of data is used for the project. Check marks show that the process has been completed, while an "x" marks that the data is not ready yet, but does note how it will be generated.

| Pilot | Open data | Synthetic data | Real data from pilot partner |
|---|---|---|---|
| Pilot 1: Medical | | ✓ | |
| Pilot 2: Mobility | ✓ | ✓ | |
| Pilot 3: Viticulture | ✓ | | ✓ |
| Pilot 4: Industry | | | ✓ |
| Pilot 5: Finance | | ✗ | |
| Pilot 6: Regional planning | ✓ | ✓ | |

*TABLE 14: OVERVIEW OF THE DATA GENERATION PROCESSES PER PILOT.*

The initial architectural framework for TEADAL, detailed in chapter 9 of this deliverable, lays a solid foundation for the further development of the project's tools. The four key architectural views—requirements, conceptual, process, and deployment—provide a structured framework that accommodates the project's diverse needs. However, further iterations are required, and the next phases expect a deeper delving into the architecture's details, ensuring it aligns with the project's objectives.

The next stages will focus on demonstrating TEADAL architecture fitness for purpose and refining design choices based on the pilots' use-cases. The respective definitions of federated data products, policies, and final analysis goals will also be targeted. The architectural work will continue to guide the solution towards a more efficient and secure digital landscape, bridging the gap between the design and practical implementation that accommodates the trustworthy, privacy-preserving, and energy-efficient data exchange needs, within federated settings, of the multiple business domains targeted in this document.



Funded by the European Union

# REFERENCES

[1] Philippe Kruchten. 1995. The 4+1 View Model of Architecture. IEEE Softw. 12, 6 (November 1995), 42–50. https://doi.org/10.1109/52.469759

[2] Z. Dehghani, Data mesh: delivering data-driven value at scale, First edition. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2022.

[3] M. N. Huhns and M. P. Singh, "Service-oriented computing: key concepts and principles," in IEEE Internet Computing, vol. 9, no. 1, pp. 75-81, Jan.-Feb. 2005, doi: 10.1109/MIC.2005.21.

[4] J. Schleier-Smith et al., "What serverless computing is and should become," Communications of the ACM, vol. 64, no. 5, pp. 76–84, May 2021, doi: 10.1145/3406011.



Funded by the European Union