CLOUD FOR DATA-DRIVEN POLICY MANAGEMENT

Project Number: 870675 Start Date of Project: 01/01/2020 Duration: 36 months

D2.7 CONCEPTUAL MODEL & REFERENCE ARCHITECTURE

Dissemination Level	PU
Due Date of Deliverable	30/6/2022 (M30)
Actual Submission Date	18/07/2022
Work Package	WP2 Requirements, Architecture & Innovation
Task	T2.2 Definition of Target Conceptual Model & Reference Architecture
Type	Report
Approval Status	
Version	v3.0
Number of Pages	p.1 – p.86

Abstract: This document provides the third version (final) of the Conceptual Model and Reference Architecture of PolicyCLOUD. It completes the work of the two previous versions of the document (Deliverable D2.2 and Deliverable D2.6). This third version, fine-tunes the definition of the overall architecture and its components, focuses on the integration of PolicyCLOUD with external frameworks, provides more information on the architecture of the Data Marketplace, highlights the internal mechanisms of the Policy Development Toolkit, presents the integration with EGI-Check-In, and describes specific enhancements based on the Legal and Ethical Framework recommendations, such as the addition of specific fields/parameters to the registration Application Programming Interfaces. This third version also addresses the Reviewers' comments for Deliverable D2.6 received within the second review report.

PolicyCloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870675.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

Versioning and Contribution History

Version	Date	Reason	Author
1.1	4/09/2020	Deliverable D2.2 Conceptual Model & Reference Architecture (submitted)	Please refer to the submitted document
2.0	30/6/2021	Deliverable D2.6 Conceptual Model & Reference Architecture (submitted)	Please refer to the submitted document
2.1	07/06/2022	Integration of External Frameworks with PolicyCLOUD (new section 7.6.11.4).	Ofer Biran (IBM), Oshrit Feder (IBM), Yosef Moatti (IBM), Nikitas Sgouros (UPRC)
2.2	08/06/2022	Overall Conceptual View and architecture of the Data Marketplace (updated section 7.9.1).	Thanos Kiourtis (UPRC), Argyro Mavrogiorgou (UPRC), George Manias (UPRC), Nikitas Marinos Sgouros (UPRC), Dimosthenis Kyriazis (UPRC)
2.3	09/06/2022	Policy Development Toolkit: User interface initialisation with Policy Model components and visualization of results (updated section 7.8.3).	Kostas Moutselos (ICCS)
2.4	10/06/2022	Update to the sections on the Ethical and Legal Compliance Framework, following up on D3.6 including description of specific fields/parameters added to the registration Application Programming Interfaces to be populated with details regarding each individual analytics tool and dataset/data source (sections 7.5.1 and 7.5.2).	Alberto Bettiol (ICTLC), Martim Taborda Barata (ICTLC)
2.5	10/06/2022	Integration of the Data Governance model, protection and privacy enforcement mechanisms with the PDT, the cloud	Konstantinos Oikonomou (UBI)

		gateways and the marketplace (section 7.10.2). Integration of EGI-Check-In with Keycloak (section 7.10.2) including the integration of Data Governance model, protection and privacy enforcement mechanisms with the Kubernetes cluster (section 7.10.2.2).	
2.6	23/6/2022	Contributions to section 8.5	Sebastian Luna-Valero (EGI)
2.7	28/6/2022	Updated section Policy Modelling & KPIs Identification (section 7.7.1)	Chris Maragkos (OKS)
2.8	6/7/2022	Peer review	Yosef Moatti (IBM), Nikos Achilleopoulos (MAG)
2.9	12/07/2022	Quality Check	Argyro Mavrogiorgou (UPRC)
3.0	18/07/2022	Final edits: Technical Coordination, Architecture Integration, Editing of document Submitted document	Panayiotis Tsanakas (ICCS), Panayiotis Michael (ICCS), Vrettos Moulos (ICCS)

Author List

Organisation	Name
ATOS	Maria Angeles Sanguino Gonzalez
ATOS	Jorge Montero Gomez
ATOS	Tomas Pariente Lobo
ATOS	Ricard Munné
EGI	Giuseppe La Rocca
EGI	Sebastian Luna-Valero
IBM	Ofer Biran
IBM	Oshrit Feder
IBM	Yosef Moatti
OKS	Chris Maragkos
ICCS	Kostas Moutselos
ICCS	Vrettos Moulos
ICCS	Panayiotis Tsanakas
ICCS	Panayiotis Michael
ICTLC	Alberto Bettiol
ICTLC	Martim Taborda Barata
ITA	Rafael del Hoyo
LON	Ebenezeer Williams
LON	Sarah Frost
LON	Adil Mohammed Ali
LXS	Jose Maria Zaragoza
LXS	Jacob Roldan
LXS	Patricio Martinez
LXS	Javier López Moratalla
LXS	Sadra Ebro
MAG	Armend Duzha
MAG	Nikos Achilleopoulos
OKS	Petya Bozhkova
OKS	Konstantinos Nasias
SARGA	Javier Sancho
SOF	Iskra Yovkova
UBI	Konstantinos Oikonomou
UBI	Giannis Ledakis
UPRC	Thanos Kiourtis

UPRC	Ilias Maglogiannis
UPRC	Argyro Mavrogiorgou
UPRC	George Manias
UPRC	Nikitas Marinos Sgouros
UPRC	Dimosthenis Kyriazis

Abbreviations and Acronyms

Abbreviation/Acronym	Definition
ABAC	Attribute-based access control
API	Application Programming Interface
CMF	Cloud Management Framework
CFREU	Charter of Fundamental Rights of the European Union
DB	Database
DPIA	Data Protection Impact Assessment
DSS	Decision Support System
EC	European Commission
ECHR	European Convention on Human Rights
EOSC	European Open Science Cloud
EU	European Union
GDPR	General Data Protection Regulation
GTD	Global Terrorism Database
IaaS	Infrastructure as a Service
JDBC	Java Database Connectivity
JSON	JavaScript Object Notation
KPI	Key Performance Indicators
ML	Machine Learning
NLP	Natural Language Processing
NoSQL	Non-Structured Query Language
OLAP	Online analytical processing
OLTP	Online transaction processing
PaaS	Platform as a Service
PDT	Policy Development Toolkit
PMO	Policy Model
PME	Policy Model Editor
PM	Policy Maker
PP	Public Policy
PR	Pattern Recognition

REST	Representational state transfer
SaaS	Software as a Service
SKA	Situational Knowledge Acquisition
SKM	Situational Knowledge Model
SOA	Service Oriented Architecture
SPA	Single Page Application
SQL	Structured Query Language
TRL	Technology Readiness Level
UI	User Interface
VM	Virtual Machine
RACI	Responsible, Accountable, Consulted and Informed model

Contents

List of Tables

List of Figures

1 Executive Summary

The third and final version of the PolicyCLOUD Conceptual Model & Reference Architecture (originally submitted as Deliverable D2.2 in September 2020 [20] with the second version submitted as D2.6 in June 2021 [21]) is presented in this document.

The PolicyCLOUD Conceptual Model presents the overall project concept along 2 main axes. Along the first data axis PolicyCLOUD delivers Cloud Gateways and APIs to access data sources and adapt to their interfaces so as to simplify interaction and data collection from any source. Along the second main axis, the Policies Management Framework of PolicyCLOUD allows the definition of forward-looking policies as well as their dynamic adaptation and refocusing to the population they are applied on. Based on the project's offerings along the main two axes of the Concept, five main building blocks (in a layered manner) define its Architecture: (1) The Cloud Based Environment and Data Acquisition, (2) Data Analytics, (3) the Policies Management Framework, (4) the Policy Development Toolkit and (5) The Marketplace. The architecture also includes a Data Governance Model, Protection and Privacy Enforcement and the Ethical Framework as depicted in Figure 2.

The architecture allows for integrated data acquisition and analytics. It also allows data fusion with processing and initial analytics (see 7.6.5) as well as seamless analytics (see 7.6.6) on hybrid data at rest. Integration in PolicyCLOUD follows three directions: (i) architecture integration, (ii) integration with the cloud infrastructure and (iii) integration with Use Case scenarios through the implementation of end-to-end scenarios. Additional integration activities take place along the two frameworks of PolicyCLOUD, (a) the Data Governance model, protection and privacy enforcement mechanism and (b) the Ethical and Legal Compliance framework. For end-to-end data path analysis we have used two Use Case scenarios: (i) the scenario of Use Case 1: "Radicalization incidents" and the scenario of Use Case 2: "Visualization of negative and positive opinions on social networks for different products".

The new updates in this final document provide the following:

- Analysis of how External Frameworks can be integrated with PolicyCLOUD (section 7.6.11.4);
- Presentation of the overall Conceptual View and architecture of the Data Marketplace (section 7.9.1);
- Outline of the mechanisms developed for initialising the Policy Development Toolkit with Policy Model components and the visualization of results (section 7.8.3);
- Analysis of the Ethical and Legal Compliance Framework positive interventions to the PolicyCLOUD architecture, including the addition of specific fields/parameters to the registration Application Programming Interfaces to be populated with details regarding each individual analytics tool and dataset/data source (section 7.5);
- Presentation of the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Policy Development Toolkit, the cloud gateways and the marketplace (section 7.10.2), and within the same context, the integration of EGI-Check-in with Keycloak including the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Kubernetes cluster.

The document also addresses the Reviewers' comments to the previous version of the deliverable (Deliverable D2.6), included in the second review report. In order to address these comments, additional updates of Deliverable D2.7 include: (i) links to specific user/stakeholder requirements (D2.5), (ii) descriptions and implementation details for the two remaining pilot Use Cases (Sofia and London) and (iii) reference to EOSC and to the role of the Conceptual Model & Reference Architecture document for the identification of the relevant services and of their providers, and description of the onboarding process based on Deliverable D3.4 [22].

2 Introduction

The definition of the Conceptual Model and Reference Architecture is a continuous, dynamically changing task, following the development of the project from M1 to M30. This document is the third (final) version of the Conceptual Model and Reference Architecture of PolicyCLOUD. The initial document has been submitted as Deliverable D2.2 [20], the second version as Deliverable D2.6 [21] and this document, Deliverable D2.7, is the final update of the document. This final update includes content from D2.2 and D2.6 , updates it where necessary, and advances it with new contributions. The new content and updates are outlined later in this introductory section and summarized in section 2.1.

The document is structured as follows: The PolicyCLOUD Conceptual Model explaining the overall project concept through 2 main axes is presented in Section 6, while the PolicyCLOUD Architecture consisting of five main building blocks (five Layers) that realize the project's offerings along the main two axes of the Concept, is presented in Section 7.

More specifically an overview of the overall architecture as presented and discussed (i) during the Kick-Off meeting, (ii) during the development of the preliminary specification as an internal report made available to partners and (iii) during specialized workshops integrating constituent architectures, is presented in section 7.2. In sections 7.3-7.9 the five layers of the architecture are presented as follows:

- **Layer 1a**-Cloud Based Environment is presented in Section 7.3.
- **Layer 1b**-Data Management – Data Stores is presented in Section 7.4.
- **Layer 2**-Data Acquisition and Analytics is presented in section 7.6.
- **Layer 3**-Policy Management Framework is presented in section 7.7.
- **Layer 4**-Policy Development Toolkit and Visualization is presented in section 7.8.
- **Layer 5**-Data Marketplace is presented in Section 7.9.

The **Ethical and Legal Compliance Framework** presented in Section 7.5 is included in the architecture from the very beginning of the project to provide extensive and in-depth analysis of relevant legal, regulatory, societal and ethical aspects.

The Data Governance Model, Protection and Privacy Enforcement used to protect data and ensure decisions across the complete path that follow specific guidelines and legislations, is presented in Section 7.10.

The architecture allows for integrated data acquisition and analytics. It also allows data fusion with processing and initial analytics (see 7.6.5) as well as seamless analytics (see 7.6.6) on hybrid data at rest.

For end-to-end data path analysis and in order to demonstrate the characteristics of the integrated architecture we have used two Use Case scenarios (section 8): The scenario of Use Case 1: "Radicalization incidents" (8.1.1) and the scenario of Use Case 2: "Visualization of negative and positive opinions on social networks for different products" (8.2.1).

Integration in PolicyCLOUD follows three directions: (i) architecture integration, (ii) integration with the cloud infrastructure and (iii) integration with Use Case scenarios through the implementation of end-to-end scenarios.

Additional integration activities take place along the two frameworks of PolicyCLOUD, (a) the Data Governance model, protection and privacy enforcement mechanism and (b) the Ethical and Legal Compliance framework.

The document analyzes how External Frameworks can be integrated with PolicyCLOUD. The overall Conceptual View and architecture of the Data Marketplace is presented, while the mechanisms developed for initialising the Policy Development Toolkit with Policy Model components and the visualization of results are outlined. Specific positive interventions to the architecture by the Ethical and Legal Compliance Framework are described. Such interventions include the addition of specific fields/parameters to the registration Application Programming Interfaces to be populated with details regarding each individual analytics tool and dataset/data source. Newly introduced in the document is also the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Policy Development Toolkit, the cloud gateways and the marketplace. Within the context of the Data Governance model, protection and privacy enforcement, the integration of EGI-Check-in with Keycloak including the integration with the Kubernetes cluster are outlined.

The new/updated sections of this document (summarized also in section 2.1) include the following:

1. Integration of External Frameworks with PolicyCLOUD (new section 7.6.11.4). The integration addresses the challenges posed by the need to integrate External Frameworks to which the serverless paradigm is not well suited.
2. The Overall Conceptual View and architecture of the Data Marketplace is presented (updated section 7.9.1).
3. Policy Development Toolkit: The mechanisms developed for the User interface initialisation with Policy Model components as also for the visualization of results is presented (updated section 7.8.3).
4. The sections referring to the Ethical and Legal Compliance Framework have been updated, to reflect additional developments and steps taken to ensure all controls identified in the developed Legal/Ethical Checklists are implemented in the Project, as reported in D3.6 [15]. Positive interventions to the PolicyCLOUD architecture, including the addition of specific fields/parameters to the registration Application Programming Interfaces, regarding each individual analytics tool and dataset/data source are also presented (sections 7.5.1 and 7.5.2).
5. Integration of the Data Governance model, protection and privacy enforcement mechanisms with the PDT, the cloud gateways and the marketplace is outlined (section 7.10.2).
6. The integration of EGI-Check-in with Keycloak including the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Kubernetes cluster are presented (section 7.10.2).

The document also addresses the Reviewers' comments for the previous version of the deliverable (Deliverable D2.6) included in the second review report. In order to address the Reviewers' comments, the following additional updates are included in Deliverable D2.7: (i) links to specific user/stakeholder requirements (D2.5), (ii) descriptions and implementation details for the two remaining pilot Use Cases (Sofia and London) - sections 8.3 and 8.4 and (iii) reference to EOSC (section 7.3.3) and to the role of the Conceptual Model & Reference Architecture document for the identification of the relevant services and of their providers, and description of the onboarding process based on Deliverable D3.4 [22].

2.1 Summary of Changes

A summary of changes is provided in the following list:

1. Integration of External Frameworks with PolicyCLOUD (new section 7.6.11.4).
2. Overall Conceptual View and architecture of the Data Marketplace (updated section 7.9.1).
3. Policy Development Toolkit: User interface initialisation with Policy Model components and visualization of results (updated section 7.8.3).
4. Update to the sections on the Ethical and Legal Compliance Framework, following up on D3.6 including description of specific fields/parameters added to the registration Application Programming Interfaces to be populated with details regarding each individual analytics tool and dataset/data source (sections 7.5.1 and 7.5.2).
5. Integration of the Data Governance model, protection and privacy enforcement mechanisms with the PDT, the cloud gateways and the marketplace (section 7.10.2).
6. Integration of EGI-Check-In with Keycloak (section 7.10.2) including the integration of Data Governance model, protection and privacy enforcement mechanisms with the Kubernetes cluster (section 7.10.2.2).
7. Links to specific user/stakeholder requirements (D2.5) (addressing Reviewers' comments).
8. Descriptions and implementation details for the two remaining pilot Use Cases, Sofia and London – sections 8.3 and 8.4 (addressing Reviewers' comments).
9. Reference to EOSC (section 7.3.3) and to the role of the Conceptual Model & Reference Architecture document for the identification of the relevant services and of their providers, and description of the onboarding process based on Deliverable D3.4 [22] (addressing Reviewers' comments).

3 Terminology

Policies KPIs are the key performance indicators (i.e. metrics/parameters) included in the structural representation of policies. These indicators are used to model the policies as well as to monitor and evaluate them.

Platform as a Service Orchestrator allows to coordinate the provisioning of virtualized compute and storage resources on Cloud Management Frameworks, both private and public (like OpenStack, OpenNebula, AWS, etc.) and the deployment of dockerized long-running services and batch jobs on Apache Mesos clusters [1].

PDT (Policy Development Toolkit) is a framework which incorporates the visualization workbench and provides a unique point of interaction with the policy makers. Through the toolkit the policy makers are able to state their questions and perform policy modelling and policy making.

Object Storage is designed to support exponential data growth and cloud-native workloads. It provides cross-region offerings, and integrated services. Depending on the access frequency of the data, storage can be provided in three **"smart tiers": Hot, Cool and Cold** [2].

External Data Sources are data sources residing at a site outside the cloud infrastructure of PolicyCLOUD (e.g. on-premise infrastructure of data owned privately by stakeholders without these being publicly released).

Interim Repository is a novel business process framework supported by a temporary storage introduced by PolicyCLOUD partners, used for caching information received from different use cases. Datasets in the interim repository are audited for their usage from a legal and ethical perspective. The Interim Repository framework enables PolicyCLOUD stakeholders to:

1. Remove data/information silos (e.g. multiple authorities from different EU countries provide their data to be used in PolicyCLOUD) ;
2. Provide the background/first stage to enable data interoperability ;
3. Provide a legal "umbrella" for the data stored in the repository (Legal and ethical assessment is performed before transferring the data to the interim repository);
4. Provide safe and authorized access to data owned by the Use Cases to the Gateways of PolicyCLOUD;
5. Provide datasets that will be used for cross-domain evidence-based policymaking;
6. Enable introductory discussions for developing scenarios based on the provided datasets, learning from the areas of interest of different use cases using the Interim Repository.

4 PolicyCLOUD offerings

PolicyCLOUD offerings are materialized through five main building blocks, supported by the **Ethical and Legal Compliance Framework** and the **Data Governance Model, Protection and Privacy Enforcement Framework**.

In summary these offerings are the following:

1. **The Cloud Capabilities & Data Collection Engine** that incorporates technologies for interfacing and acquiring data from various sources.

2. **The Reusable Models & Analytical Tools Engine** that incorporates all data services / technologies provided by PolicyCLOUD for the data path/lifecycle.

3. **The Policies Management Framework**.

4. **The Policy Development Toolkit** providing an interactive environment and the Front-End of the system.

5. **The Data Marketplace** which enables data and knowledge to be exploited as assets while keeping conformance with legal and ethical requirements and privacy protection.

Finally, the **Ethical and Legal Compliance Framework** has been designed to identify and monitor the implementation of measures to address relevant ethical, legal and security aspects applicable to all PolicyCLOUD offerings, while the **Data Governance Model, Protection and Privacy Enforcement Framework** protects data and ensures decisions across the complete path following specific guidelines and legislations.

The details of the PolicyCLOUD offerings listed above are provided in section 7.1 with title Architecture Building Blocks.

5 PolicyCLOUD capabilities

PolicyCLOUD provides an innovative suite of state-of-the-art technology capabilities and management frameworks over a Cloud environment as presented in the following list:

- **Cloud Based environment** to support the development of PolicyCLOUD using Platform as a Service (PaaS) solutions.
- **Unified Cloud Gateway** moving streaming and batch data from data owners into PolicyCLOUD layers while performing data source reliability.
- **Incentives' identification and management** offering a set of tools to identify and manage incentives able to engage different participants on the policy making process.
- **Access to heterogeneous data stores.**
- **Scalable Database** with the ability to scale out over hundreds of nodes.
- **Polyglot capabilities** enabling the Querying of Heterogeneous Data Sources in a Unified manner.
- **Ability to combine analytics on streaming data and on data at rest.**
- **Transparent to the user movement of colder data** to the Object Store tier.
- **Data Cleaning** for the detection and correction of corrupted or inaccurate records received from Cloud Gateways.
- **Data Interoperability** based on data-driven design, coupled with linked data technologies, in order to improve both semantic and syntactic data and dataset interoperability.
- **A business process for clearing private data, as well as "open data"** evaluating if and to which extent personal data in terms of the GDPR is allowed to be processed by PolicyCLOUD.
- **Data Fusion** task permitting the merging of data coming from disparate sources into a single data set, integrated with initial analytics and data processing tasks.
- **Seamless Analytics** permitting undifferentiated access and query capabilities both to hot (in the DB) and cold (in the object storage) data.
- **Situational knowledge** from data from sensors, social media and datasets offering feature extraction, clustering and categorization.
- **Opinion Mining** providing social attitude regarding specific topics, identifying specific entities and generating a “contributor graph” based on discussions of various policies from citizens.
- **Sentiment analysis** based on the input received from the pilots about their policies.
- **Social Dynamics** providing a concurrent, web-based environment for social simulation. The environment allows users to create graph-based population models online. The presented Social Dynamics framework is actionable from the PolicyCLOUD environment.

- **Framework for Cloud usage by Public Authorities** examining (a) the different mechanisms, methods and technologies used for policy lifecycle and (b) a proposition of a set of adaptable techniques towards the utilization of cloud environments for policies creation.
- **Middleware for modelling and designing of Policies** providing a mechanism for policies to be modelled and designed based on specific structural representations, allowing users to create a policy by selecting a schema of data, applying well known Key Performance Indicators.
- **Policy Development Toolkit (PDT)** constituting the Front-End of PolicyCLOUD environment. It integrates several sub-components to enable policy makers to create, update and validate policy models.
- **Integrated cloud-based framework** designed for the Cloud, structured over five layers including an Ethical and Legal Compliance Framework and a Data Governance Model providing all the above capabilities.

6 PolicyCLOUD Conceptual Model

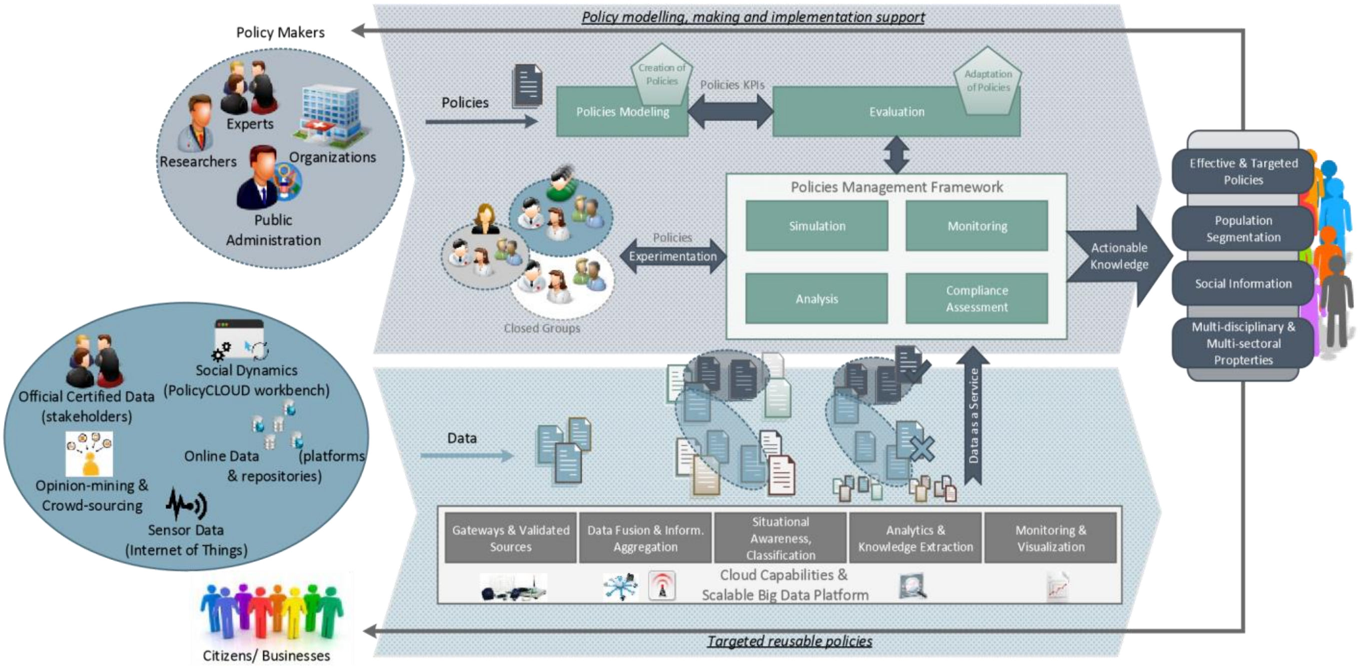

Policy modelling, making and implementation support
Policy Makers
Experts
Researchers
Organizations
Public Administration
Policies
Creation of Policies
Policies Modeling
Policies KPIs
Adaptation of Policies
Evaluation
Policies Management Framework
Simulation
Monitoring
Analysis
Compliance Assessment
Policies Experimentation
Closed Groups
Actionable Knowledge
Effective & Targeted Policies
Population Segmentation
Social Information
Multi-disciplinary & Multi-sectoral Propterties
Official Certified Data (stakeholders)
Social Dynamics (PolicyCLOUD workbench)
Online Data (platforms & repositories)
Opinion-mining & Crowd-sourcing
Sensor Data (Internet of Things)
Data
Data as a Service
Gateways & Validated Sources
Data Fusion & Inform. Aggregation
Situational Awareness, Classification
Analytics & Knowledge Extraction
Monitoring & Visualization
Cloud Capabilities & Scalable Big Data Platform
Citizens/ Businesses
Targeted reusable policies

FIGURE 1 – THE POLICYCLOUD CONCEPTUAL MODEL

PolicyCLOUD architecture delivers a set of innovative technologies with an overall goal to enable data-driven management of policies lifecycle, from their modelling and implementation, to optimization, compliance monitoring, adaptation and enforcement.

As depicted Figure 1, PolicyCLOUD architecture enables the compilation of multi-disciplinary, and multi-sectoral optimized policies. Multi-disciplinary policies aim at addressing different spatiotemporal levels. In terms of time scales, different policies are proposed to be applied in long-term, while these policies could address a specific area (e.g. city), a region, or even a country. The combination of these properties of policies are optimized through PolicyCLOUD according to the modelling and evaluation of different policies and their corresponding KPIs.

Additionally, data emerging from policies "collections" / clusters (e.g. all policies in a city, environmental policies in different cities, health policies for specific age groups, etc) provide additional information for the optimization of policies in the aforementioned scale. Furthermore, PolicyCLOUD architecture enable multi-sectoral optimization of policies.

As shown in Figure 1, policies effectiveness is assessed and optimized based on their KPIs (vertical optimization) while KPIs of policies from other sectors are also taken into consideration (horizontal, cross-sector optimization). To realize the overall multi-sectoral effectiveness of policies, PolicyCLOUD architecture includes technologies for correlation of the policies and the data used to compile these policies through reusable and scalable models and analytic tools.

The architecture serves the overall project concept of PolicyCLOUD and it is realized through 2 main axes: the data axis and the policies axis (Figure 1).

Along the first data axis PolicyCLOUD delivers Cloud Gateways and APIs to model the data sources and adapt to their interfaces so as to simplify interaction and data collection from many different sources.

Some of these sources may not provide reliable information and thus before taking it into consideration, gateways are enhanced with the functionality of validating the data in order, with appropriate business processes, to develop trust and reliability profiles and patterns of the sources and exploit only the reliable ones.

In terms of data sources, PolicyCLOUD obtains open data from the ecosystem stakeholders (e.g. public authorities), sensor data from Internet of Things infrastructures (e.g. environmental sensors), data from online platforms, opinion-mining and crowd-sourcing data (both from online platforms and from the proposed PolicyCLOUD living lab approach), as well as data related to social dynamics and behaviour through the corresponding analytical tools.

The Ethical and Legal Compliance Framework included in the architecture enables a process we name "data clearance" which examines available open-data for privacy issues (even if some data are characterized as "open" they may include private data) and other relevant legal and ethical concerns (e.g., databases subject to specific licensing terms, potential for bias or misrepresentations within a dataset). Data clearance processing combines legal/ethical expertise with technology (e.g. access control at critical points) in order to safeguard that data are efficiently used in a legal and ethical manner.

Based on the above, data fusion and information aggregation enable the compilation of information into new data and metadata structures which are interlinked and analyzed. This information along with existing policies provide a network of knowledge which is dynamically exploited for improving the effectiveness of existing policies and facilitating the creation and adoption of new policies.

PolicyCLOUD architecture delivers mechanisms for clustering, classification and situational awareness on big datasets and the corresponding policies. Core element in this process is the delivery of a powerful Reusable Models & Analytical Tools offering for cleaning datasets, modelling and representing them, as well as harnessing information and enabling knowledge extraction. This is performed by taking into consideration data and existing policies that correspond to target groups / public authorities with specific goals and population characteristics.

Given the wealth of information and the different administrative and legal domains under which data will be governed and managed, PolicyCLOUD includes a data governance model (based on RACI) that governs the complete data lifecycle (e.g. who has access, to which data, etc).

Along the second main axis the Policies Management Framework of PolicyCLOUD is exploited for the definition of forward-looking policies which are dynamically adapted and methodically focused on the population that are applied on.

Initially the policies are modelled in order to extract quantitative and qualitative information from them, such as KPIs, operational and functional dependencies for analysis and evaluation.

The architectural framework employs the knowledge incorporated into the clusters of data and policies for a) assessing and stratifying the risks of policies, b) monitoring and assessing their compliance and c) forecasting the effectiveness of policies, including variations and combinations of policies.

The process is supported both by simulation methodologies and techniques, as well as by analysing the results of applying the policies to closed groups – i.e. evidence-based. Evaluation is not based on policy-level but on KPI-level per policy and across sectors (addressing different verticals including environment, migration, employment, etc.). In addition, through the mechanisms described in the architecture, the policies strengths and weaknesses are identified and analyzed while when it comes to policies adoption, their effectiveness on different conditions, populations, methodologies etc. is effectively assessed.

Therefore, the policies not only are evaluated, but they are also differentiated with different parameter sets, applicable to certain groups, locations and conditions, with in advance knowledge of the risk and performance trade-offs. Identification of the exact elements of policies that can affect their outcomes, across all policies, will also enable the creation of policies taking advantage of the excellence of the particular elements on better and more targeted results, minimizing in parallel the uncertainty when integrating them in the public policy strategy.

The outcomes - as actionable knowledge - are delivered to policy makers as evidence-based targeted strategies for policy making (including the most relevant population segmentation and evidences to maximize the policies efficiency).

The Conceptual Model has been revisited after the release of the second version of this document (Deliverable D2.6). During an internal workshop with all PolicyCLOUD partners, coordinated by ICCS, the Conceptual Model was re-examined based on the acquired knowledge and experience received while integrating the use-case scenarios and while interacting with end-users during the co-creation workshops. The internal workshop re-validated the Conceptual Model which is included in this final version of the document (Deliverable D2.7) without the need of any amendments.

7 PolicyCLOUD Architecture

7.1 Architecture Building Blocks

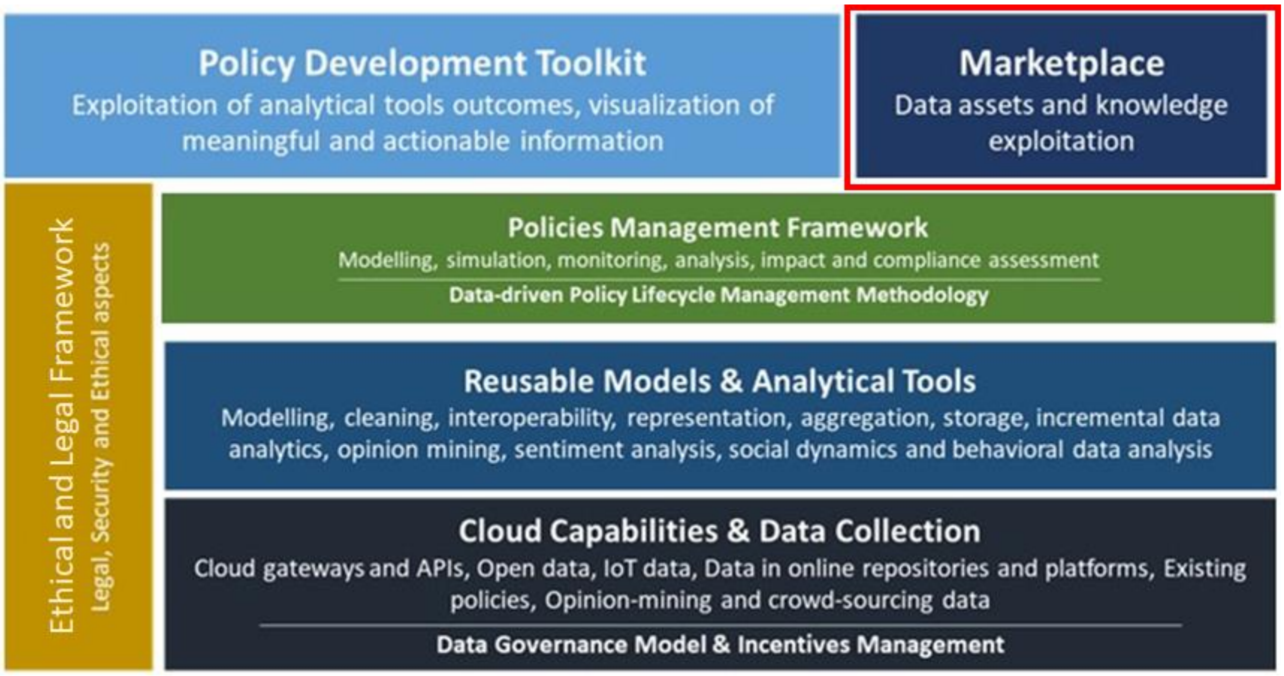

Policy Development Toolkit
Exploitation of analytical tools outcomes, visualization of meaningful and actionable information
Marketplace
Data assets and knowledge exploitation
Ethical and Legal Framework
Legal, Security and Ethical aspects
Policies Management Framework
Modelling, simulation, monitoring, analysis, impact and compliance assessment
Data-driven Policy Lifecycle Management Methodology
Reusable Models & Analytical Tools
Modelling, cleaning, interoperability, representation, aggregation, storage, incremental data analytics, opinion mining, sentiment analysis, social dynamics and behavioral data analysis
Cloud Capabilities & Data Collection
Cloud gateways and APIs, Open data, IoT data, Data in online repositories and platforms, Existing policies, Opinion-mining and crowd-sourcing data
Data Governance Model & Incentives Management

FIGURE 2 – POLICYCLOUD ARCHITECTURE BUILDING BLOCKS

The architecture of PolicyCLOUD includes five main building blocks that realize the project's offerings (Figure 2) along the main two axes of the Concept described in the previous section. These building blocks are presented in the following paragraphs in a bottom-up manner:

1. The **first building block** of the PolicyCLOUD architecture is the **Cloud Capabilities & Data Collection Engine** block that incorporates technologies for interfacing and acquiring data from different sources (through unified cloud gateways and APIs), assessing their reliability and attaching the corresponding metadata to the sources and ensuring privacy enforcement for the collected data, using the developed cloud infrastructure management. This block also includes mechanisms for identifying attributes of data and stakeholders in order to ensure that all data decisions are according to the data governance rules specified by the data owners, while it integrates techniques for managing the incentives in order to ensure citizens participation.
2. The **second building block** of the architecture is the **Reusable Models & Analytical Tools Engine** that incorporates all data services / technologies provided by PolicyCLOUD for the data path / lifecycle: modelling, cleaning, interoperability, linking / aggregation, storage and incremental analytics, for constructing the required reusable models. Moreover, this engine will also offer techniques for sentiment analysis from different online platforms, and tools for opinion-mining

allowing stakeholders to "develop" through the provided toolkit, in an automated way, different means (such as aspect ranking) in order to acquire and analyse the corresponding information from citizens.

3. The **third building block** refers to the **Policies Management Framework** that incorporates services for the identification of the required KPIs in order to model the policies and identify potential interdependencies with other policies within and across sectors at different levels (section 7.7.1). The framework also includes tools for collecting evidence monitoring information both from the engaged citizens and from the population targeted by the policies, while also assessing the compliance to these policies and thus assessing the policies impact (based on the identified KPIs).
4. The **fourth building block** (the interactive environment) provides the **Policy Development Toolkit** allowing policy makers to interact with the models and analytical tools as well as to specify their requirements and constraints with respect to different policies. In addition, the toolkit facilitates visualization of policies monitoring in an adaptive and incremental way.
5. The **fifth building block** of the architecture is the **Data Marketplace** which enables data and knowledge to be exploited as assets. Data Marketplace has two goals: (a) the usage of data in different contexts (scenarios for policy making) and (b) the identification of market opportunities. Data Marketplace (section 7.9.1) is related to the overall PolicyCLOUD environment through its many different APIs able to store several types of PolicyCLOUD assets (solutions). The Data Marketplace is a standalone platform which can be integrated in the future (after the end of the project) with other systems/environments based on its own architecture presented in Figure 20. For this reason Data Marketplace is outlined in both the diagrams of Figure 2 and Figure 4 with a red rectangle.

The **Ethical and Legal Compliance Framework** has been designed to identify and monitor the implementation of measures to address relevant ethical, legal and security aspects applicable to all PolicyCLOUD offerings, thus ensuring the sustainability of the modelled policies.

The architecture building blocks have been implemented over the European Cloud Initiative infrastructure offered by EGI (Figure 3). For the research purposes of the project, PolicyCLOUD partners have successfully developed a serverless environment on top of the IaaS-type cloud provisioned through EGI (Section 7.6.11 Data Acquisition and Analytics Integration). Based on the experimental work performed, it is realized that a commercial deployment of PolicyCLOUD should be done above a serverless (or FaaS) framework.

The PolicyCLOUD Marketplace is part of the infrastructure and offers the solutions in terms of models and analytical tools that can be exploited by the end-users (i.e. policy makers and public authorities) through the PolicyCLOUD Policy Development Toolkit.

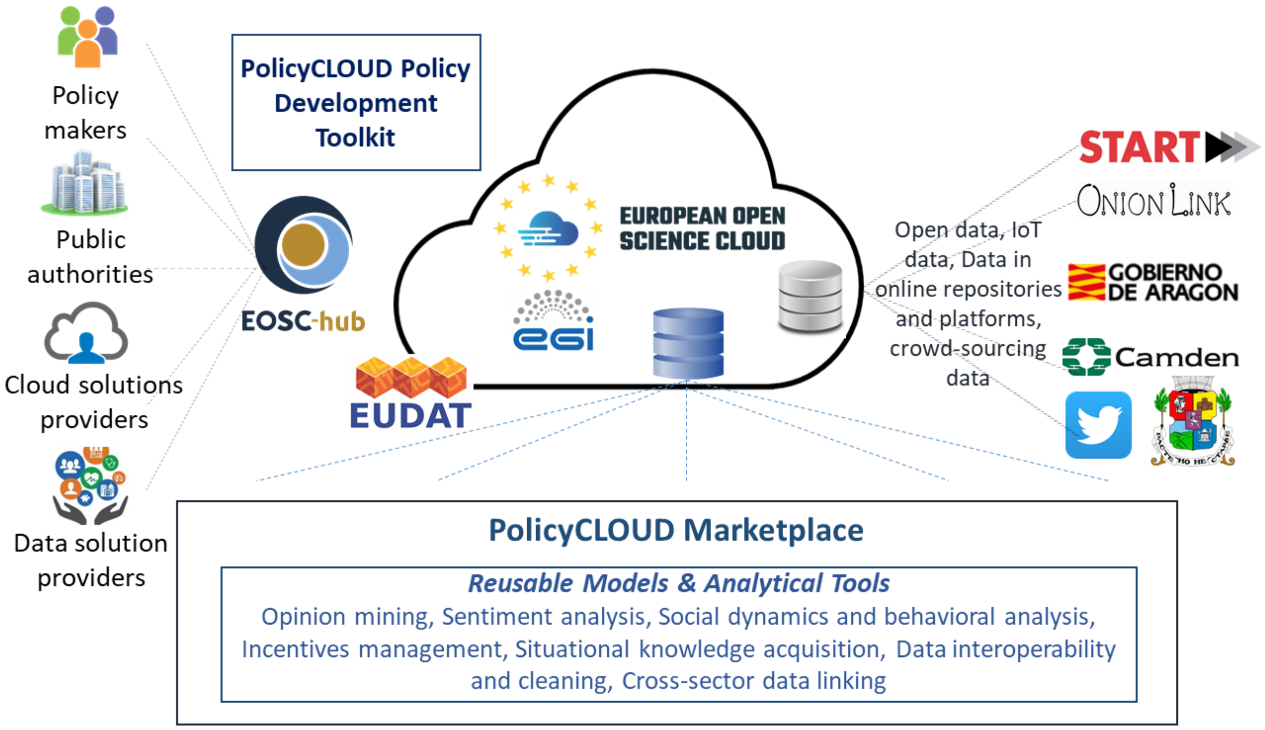

Policy makers
Public authorities
Cloud solutions providers
Data solution providers
PolicyCLOUD Policy Development Toolkit
EOSC-hub
EUDAT
EUROPEAN OPEN SCIENCE CLOUD
egi
Open data, IoT data, Data in online repositories and platforms, crowd-sourcing data
START
ONION LINK
GOBIERNO DE ARAGON
Camden
PolicyCLOUD Marketplace
Reusable Models & Analytical Tools
Opinion mining, Sentiment analysis, Social dynamics and behavioral analysis, Incentives management, Situational knowledge acquisition, Data interoperability and cleaning, Cross-sector data linking

FIGURE 3 – POLICYCLOUD ARCHITECTURE IMPLEMENTATION OVER THE EUROPEAN CLOUD INITIATIVE INFRASTRUCTURE OFFERED BY EGI

7.2 Architecture Overview

The Overall Architecture (Figure 4) has been discussed and further developed (i) during the Kick-Off meeting, (ii) during the development of the preliminary specification as an internal report made available to partners and (iii) during specialized workshops integrating constituent architectures. The architecture's layers and frameworks will be analyzed in the sections that follow.

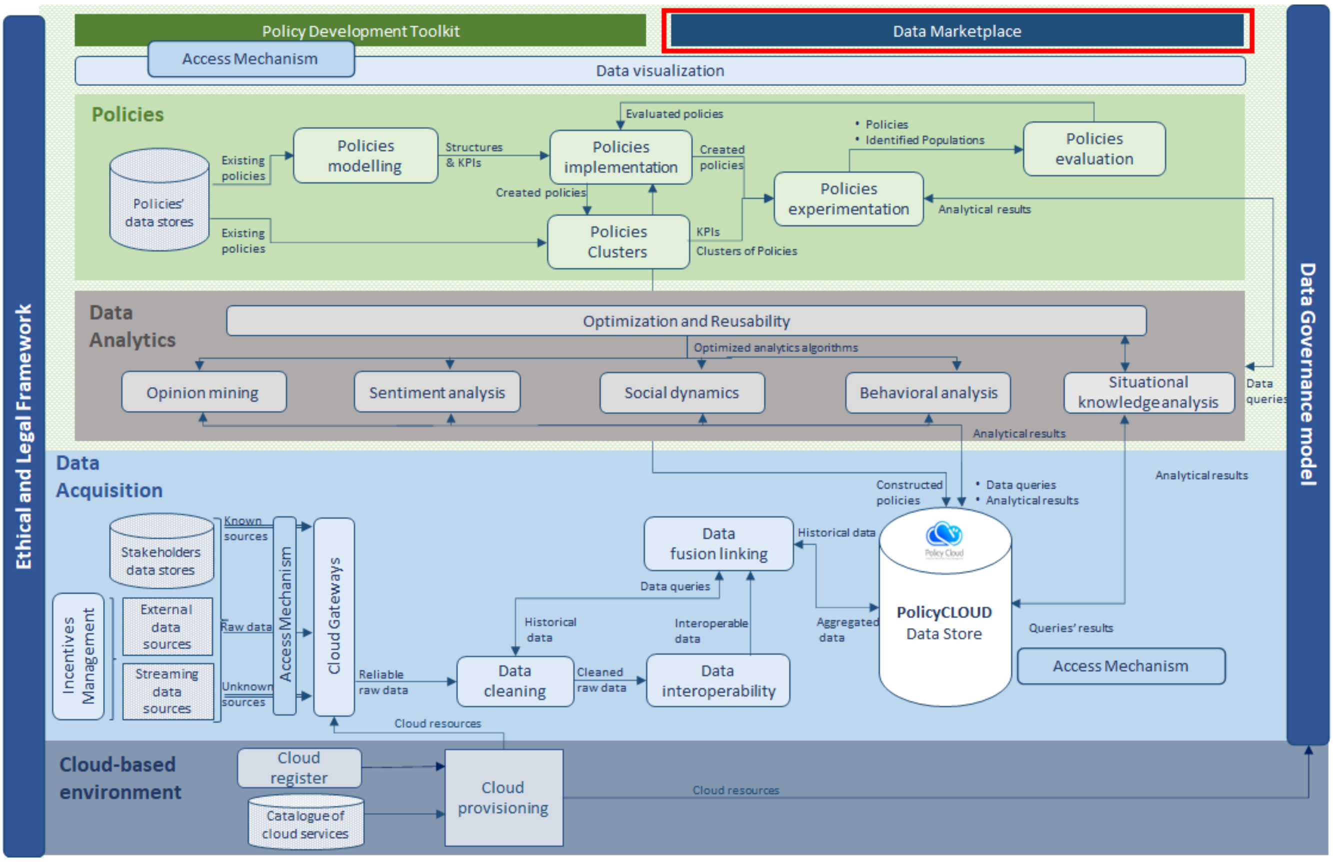


FIGURE 4 – POLICYCLOUD OVERALL ARCHITECTURE

As a complete environment, the proposed architectural approach is presented in Figure 4. The overall flow is initiated from various data sources, as depicted in the figure through the respective *Data Acquisition* block. Data sources can be data stores from public authorities or external data sources that contribute data following the provision of incentives, facilitated through the *incentives management* mechanism.

A set of APIs incorporated in a gateway component, enable data collection by applying techniques to identify the reliable sources and for these sources obtain the data and perform the required *data quality assessment and cleaning*. *Semantic and syntactic interoperability* techniques are utilized over the cleaned data providing the respective interoperable datasets to the PolicyCLOUD datastore following the required *data linking and aggregation* processes.

The datastore is accessible from a set of machine learning models represented through the *Data Analytics* building block. Machine learning models may incorporate opinion mining, sentiment and social dynamic analysis, behavioural analysis and situational / context knowledge acquisition. The data store and the analytics models are hosted and executed in a *cloud-based environment*. For this purpose, a catalogue in which an extensible set of services are registered has been implemented. Furthermore, all the analytics models are realized as services, thus enabling their invocation through a proposed policy development toolkit – realized in the scope of the *Policies* building block of the proposed architecture as a single point of entry into the PolicyCLOUD platform.

The toolkit allows the compilation of *policies as data models*, i.e. structural representations that include key performance indicators (KPIs) as a means to set specific parameters (and their target values) and monitor the implementation and effectiveness of policies against these KPIs along with the list of analytical tools to be used for their computation. According to these analytics outcomes, the values of the KPIs are specified resulting to *policies implementation / creation*. It should be noted that PolicyCLOUD also introduces the concept of *policies clusters* (section 7.7.1) in order to interlink different policies, and identify the KPIs and parameters that can be optimized in such policy collections.

Across the complete environment, an implemented *data governance and compliance model* is enforced, ranging from the provision of cloud resources regarding the storage and analysis of data to the management of policies across their lifecycle.

As shown in Figure 4, the architecture is ready for accessing External Data sources having the following characteristics:

1. External Data Sources which are not eligible to be physically imported to the central persistent storage of the platform remaining on premise due to data regulator constraints or due to excessive ingestion/maintenance costs.
2. External Data which may not be owned by the organization and thus they cannot be retrieved and ingested to the platform, or they cannot be ingested due to privacy considerations, or they might be owned by the organization, but cannot be imported due to technological constraints, and thus they are considered as external to the platform.

We need to mention that during the implementation of the PolicyCLOUD environment, accessing of External Data sources was not pursued after a series of internal meetings.

The main reason of this decision is that external access was not required by any of the PolicyCLOUD Use Cases and insisting on such implementation would require much longer time than initially expected, in order to resolve technical and regulatory issues, creating significant delays to the project with the risk of defocusing it from its aims.

Despite the fact that External Data sources have not been demonstrated with a specific Use Case scenario, as mentioned earlier, the architecture supports accessibility of External Data sources as shown in Figure 4, a feature which could be further developed and exploited when it will be required by additional Use Cases after completion of the project and during its commercialisation phase.

In summary, the architecture is ready to support accessibility of External Data sources when this is required by specific Use Cases/scenarios in the future.

7.3 Layer 1a - Cloud Based Environment

7.3.1 The EGI Federated Cloud

The EGI Federated Cloud is an IaaS-type cloud, made of academic private clouds and virtualized resources and built around open standards. Its development is driven by requirements of the scientific community. The Federation pools services from a heterogeneous set of cloud providers using a single authentication and authorisation framework that allows the portability of workloads across multiple providers and enables bringing computing to data. The current implementation is focused on IaaS services but can be easily applied to PaaS and SaaS layers. The EGI Federated Cloud architecture is based on the concept of an abstract Cloud Management Framework (CMF) that supports a set of cloud interfaces to communities.

Each resource centre of the infrastructure operates an instance of this CMF according to its own technology preferences and integrates it with the federation by interacting with EGI core components:

- Service registry for configuration management of federated cloud services.
- EGI AAI for authentication and authorisation across the whole cloud federation.
- Accounting for collecting, and displaying usage information.
- Information discovery about capabilities and services available in the federation.
- Virtual Machine image catalogue and distribution, replicating VM images as needed by the user communities in a secure way.
- Monitoring, performing service availability monitoring and reporting of the distributed cloud service endpoints.

Users of the EGI Federated Cloud infrastructure can interact with cloud providers in several ways:

- Directly using the IaaS APIs of the resource centres to manage individual resources.
- Leveraging federated IaaS provisioning tools that allow managing and combining resources from different providers enabling the portability of application deployments between them. The EGI Federated Cloud task force is currently in the process of evaluating and selecting the best tools for this task.
- Using PaaS solutions such as the Infrastructure Manager (IM)[1], a Federated IaaS Provisioning tool, or the PaaS orchestrator developed within INDIGO-DataCloud[2].

In the context of the PolicyCLOUD project, EGI contributes to the provisioning of the needed computing resources to set-up the PolicyCLOUD infrastructure. This cloud infrastructure will help policy makers,

[1] See: https://www.grycap.upv.es/im/index.php
[2] See: https://www.indigo-datacloud.eu/

public authorities and different stakeholders, to analyse a plethora of datasets from different data sources, and facilitate policy making. EGI offering for the project includes a federated IaaS cloud to run compute - or data -intensive tasks and host online services in virtual machines or Docker containers on IT resources accessible via a uniform interface. More details about the federated EGI Cloud infrastructure and the solutions offered to address the needs of the project have been provided in D3.1 - Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 1. The Requirements for Cloud Capabilities and Data Collection and Cloud Provisioning are provided in Deliverable D2.5 [23] (section 5.1 and section 6.1).

7.3.2 Integration enabling cloud infrastructure

Considering that most of the components of the PolicyCLOUD infrastructure are dockerized and distributed as Docker containers, to facilitate the provisioning of compute and storage resources, and the orchestration of distributed Kubernetes clusters, the access to the cloud infrastructure is also enabled via the PaaS orchestrator developed in the context of the INDIGO-DataCloud project[3]. For more details about the main architecture of the INDIGO-Data Cloud PaaS Orchestrator[4], please refer to deliverable D3.2 "Cloud Infrastructure Incentives Management and Data Governance Software Prototype 1". A trial phase was planned in August 2020, during which the 10-20% of the full capacity of the PolicyCLOUD infrastructure was configured to allow technical partners to run tests and assess its performance. More specifically, during this phase, LXS, IBM and ICCS managed to test the deployment of a GitLab instance for the project and the deployment of OpenWhisk on Kubernetes. In October 2020 the project signed a pay-for-use agreement with the cloud resource provider (RECAS-BARI) and the full needed capacity was allocated to the project. Resources are scaled-up on-demand with latest increase, the increase of the allocated vCPUs by 30% (May 2022).

7.3.3 Registration of PolicyCLOUD services in EOSC portal

The European Open Science Cloud (EOSC) is an environment for hosting and processing research data to support EU science. Registration of PolicyCLOUD services in EOSC portal[5] is a goal which requires preparation and assessment work, on the basis of a clear alignment on the approach to these services. EGI approaches this activity within the Consortium.

Within this context, analysis of PolicyCLOUD service maturity and gap analysis are performed. Deliverable D2.7 'Conceptual Model & Reference Architecture', consists of a reference for the identification of the relevant services and their providers (Consortium partners) while establishing a basis of discussion for the maturity of these services.

[3] https://cordis.europa.eu/project/id/653549
[4] https://indigo-paas.cloud.ba.infn.it/home/login
[5] https://eosc-portal.eu/

As analyzed in Deliverable D3.4 Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 2 [22], EGI approaches the onboarding activity through the following phases:

1. Analysis of providers who are part of the Consortium and preparation for onboarding;
2. Onboarding of providers;
3. Analysis of PolicyCLOUD service maturity and gap analysis;
4. Provision of support to onboard the services, based on the following requirements:
    a. The service is accessible to users outside its original community;
    b. The service is described through a common template focused on value proposition and functional capabilities;
    c. At least one service instance is running in a production environment available to the user community;
    d. Published research data is Findable, Accessible, Interoperable and Reusable;
    e. Release notes and sufficient documentation are available;
    f. Helpdesk channels are available for support, bug reporting and requirements gathering;
5. Technical integration;
6. Assessment of any gaps.

The process is ongoing and more information will be provided in the next version of the Deliverable: D3.7 Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 3.

7.4 Layer 1b - Data Management and Data Stores

Components: Cloud Gateways (T3.3), Incentives Management (T3.4), Data Store (Figure 4).

7.4.1 Cloud Gateways

In the context of PolicyCLOUD, the Cloud Gateway and API component developed by UPRC seeks to enhance the abilities and services offered by a unified Gateway to move streaming and batch data from data owners into PolicyCLOUD data stores layers, which support both SQL and NoSQL data stores and public and private data. Based on the specifications provided in D3.1 Cloud Infrastructure Incentives Management and Data Governance Design and Open Specification 1 [3] of the PolicyCLOUD project, the effort related to Cloud Gateways & APIs component focuses on providing a complete and “smart” entryway into PolicyCLOUD project, allowing multiple APIs or microservices to act cohesively and thus provide a uniform, gratifying experience to each stakeholder. The provided Gateway API allows building scalable and robust APIs, while simplifying the interaction and data collection from various sources and providers. The main goal of this component is to handle a request by invoking multiple microservices and aggregating the results. Hence, it enhances the design of resources and structure, add dynamic routing parameters and develop custom authorizations logic. PolicyCLOUD’s Cloud Gateway and API component supports scalability, high availability and shared state without compromising performance.

Moreover, it supports client side load balancing, so that the overall system can apply complex balancing strategies and do caching, batching, fault tolerance, service discovery and handle multiple protocols. To this end, MoleculerJS [6], a framework that bases its functionality on microservices architecture methodology, is being utilized as the core element of Cloud Gateway component. MoleculerJS framework has built-in microservices that can support the above characteristics, such as load balancing [4] or fault tolerance [5]. The latter is also being addressed though the integration with the Kafka [6] event streaming platform, one of the main tools utilized in the PolicyCLOUD project and which is used as a buffering mechanism and a message bus for providing and moving data across the whole data pipeline and across all different analytical components.

Through this ability the gateway is also able to directly ingest incoming data into the appropriate data store based on their privacy level. This feature is available for implementation on Use Cases/scenarios where direct ingestion is favoured. We need to mention that for the Use Cases/scenarios we have worked with during the project the preferred methodology to be used was through the Interim Repository (section 7.5.1) and not through direct ingestion. Nevertheless, the architecture is ready to support direct ingestion for new scenarios in the future after the completion of the project. Therefore, it makes easy to differentiate the queries/requests having to be redirected to the overall data management, analysis and storage system of the project. On top of all these, through appropriate business processes we identify the reliability levels of both all the available data sources and their incoming data, thus "feeding" into the PolicyCLOUD platform only the reliable data that comes from only reliable data sources. In this context, the Gateway is able to map all the incoming data sources to specific levels of trustfulness, and thus capturing their reliability. As a result, all the data sources that do not meet the trustfulness criteria are excluded, ensuring the origination of the data sources' incoming data, the adaptive selection of all these available data sources in order to be kept connected into the PolicyCLOUD platform. The component ensures also that the collection of the data comes only from reliable data sources so as to be used for further analysis. Furthermore, in terms of integration with other internal components and mechanisms of the PolicyCLOUD platform, the Cloud Gateway & APIs component has been successfully integrated with the Access Mechanisms in order to ensure that all the required security standards are being met and that specific roles and privileges are being defined precisely. On top of this, the Cloud Gateway & APIs component has also been integrated with OpenWhisk tool which is utilized in the scope of PolicyCLOUD platform in order to provide a serverless, holistic, integrated and end-to-end pipeline of the Data Acquisition and Analytics layer. Finally, the gateway includes an API documentation page, by using Swagger UI, in terms of providing a graphical interface for interacting with the API. The latter facilitates the exploration of all available requests and responses that are listed including also the required parameters.

[6] https://moleculer.services/

7.4.2 Incentives Management

The overall idea of Incentives Management is to offer a set of tools to identify, declare, track and manage incentives activities to engage the different participants on the policy making process, understanding their motivations in the light of the context. Therefore, this task will provide tools to manage the incentives activities performed with the policies stakeholders, either through closed groups, like some communities evaluating some proposed policies, or even engaged citizens.

The different incentives may be of different types, social, cultural, political or other types. For this purpose, the Incentives Management component may provide to the policy maker access to results from policies on PolicyCLOUD in order to create and manage incentives that relate to these results.

Similarly, the component will manage and keep track of the incentive actions proposed by the policy maker in order to involve the participants and evaluate the outcomes of these actions.

More specifically, and from a theoretical point of view, Incentives Management activities pursue to provide an individual incentives plan that will define a set of rewards corresponding to specific participant actions.

Following the four dimensions introduced by [7] (Malone, 2010): what, who, why, and how, the incentives plan will be pre-established as follows:

- **WHO (participants/requesters):** The Incentives Management task will be focused on engaging citizens, organizations who may be affected by the introduction of policies defined in PolicyCLOUD. In the case of PolicyCLOUD the exact group of citizens and/or organizations will be settled attending the existing use cases and drive by the policy maker.
- **WHAT (actions/tasks):** Is the information exchange, contributions and collaboration expected by the participants.
- **HOW (way or manner):** Define how the participants collaboration is expected. In the case of PolicyCLOUD, the way of collaboration will be established in the context of the existing use cases and drive by the policy maker.
- **WHY (rewards/incentives):** It is aimed to the establishment of different types of incentives (e.g. social, cultural, political, etc.) in return for the participant collaborations done through the execution of existing tasks (what) performed in a concreted way (how). In the case of PolicyCLOUD, the incentives will be established in the context of the existing use cases and drive by the policy maker.

Citing the description included in D3.1 deliverable, the Incentives Management activity will be focused on the following: *"Provide the maximum support to the policy maker… toward a twofold aim: support the policy maker in the incentives identification and help the policy maker in the incentives management".* [3]

7.4.2.1 Incentives Management Architecture Integration

The tool will provide policy makers the possibility to declare and track incentives actions. As pictured in Figure 5, the integration point of the Incentives component is the PDT interface where the component frontend will be integrated as an additional entry point for the policy makers, so as to have all the needed components accessible from the same access point. It may also be possible to show to the policy maker information from the policy models' KPIs they have already declared in order to better shape and adapt the incentives actions.

Crowdsourcing tools may be used by the policy maker, but those will be kept totally separate from the Incentives Management component or any other components or modules from PolicyCLOUD. It will be on policy maker discretional use of the results and information gathered through these Crowdsourcing tools that they may shape and track specific incentive actions with their corresponding policy stakeholders.

As per the backend, the component will be managing the access to the different entities with a corresponding data storage tied to it. For more details, please refer to next Deliverable D3.4 "Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 2" [8].

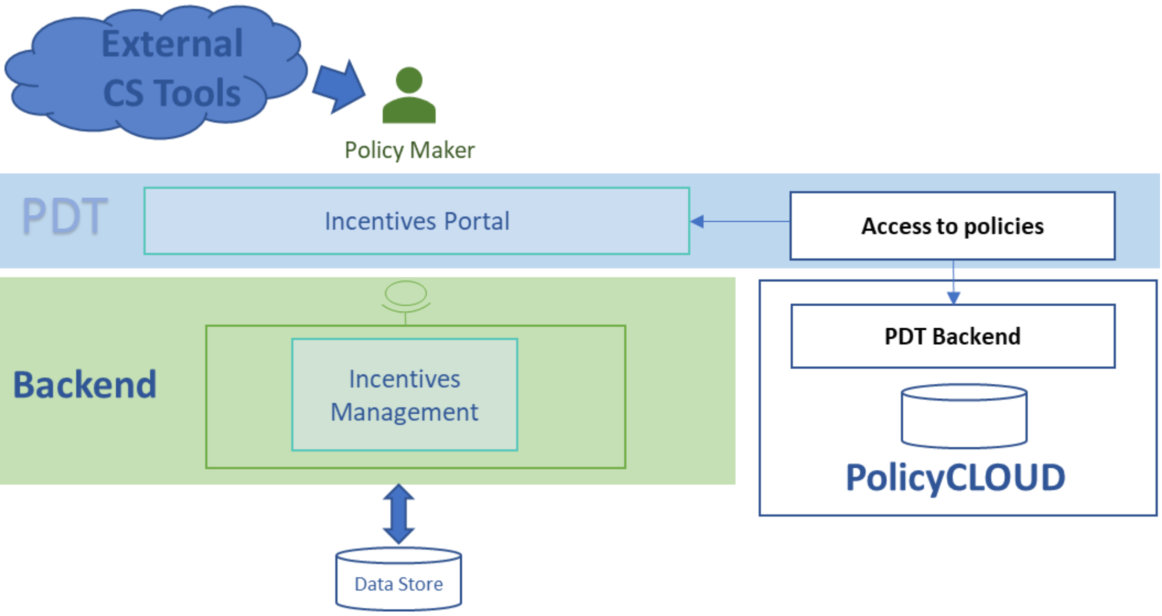


FIGURE 5 – INCENTIVES MANAGEMENT ARCHITECTURE INTEGRATION

7.4.3 Data Management and Data Stores

In the scope of the PolicyCLOUD project, different challenges are being raised regarding data management, an internal part of the data acquisition process, as data stored into the data repository of the platform are being accessed in different and heterogeneous manners. Firstly, as part of the project itself, multiple scenarios from different use case providers have been integrated to the common platform. At the same time, the platform itself is envisioned to be exploited in the future by other cases. Each one of these independent organizations (from the four use cases) is currently using its own data management systems, relying on different types of data schemas, while there is need for a central data repository to fit the needs of all. Secondly, each organization typically has different silos, relying on heterogeneous data stores for data persistency, using completely different data models: from traditional relational database systems, NoSQL databases, Hadoop datalakes etc. Moreover, the PolicyCLOUD vision is to deal with different in nature data, that is, data at-rest which typically refers to data that is permanently stored and various queries are being executed in order to retrieve the results, and streaming data that refers to data that are being continuously inserted to the system without always the need for persistent storage, but with the ability to apply automatic analytics on top of them. According to the requirements defined in Deliverable D2.1, there is the need for support of hybrid workloads, such as OLTP workloads for managing operational data and ensuring transactional semantics, and OLAP workloads in order to perform analytical queries over the operational data, while ensuring the data consistency. Finally, as operational data usually become obsolete after a certain point in time with rare modifications and in order to cope with analytics over big data, the data are transferred to Object Storage, which is much less expensive and also has infinite scalability more suitable for performing this type of analytics. The requirement in this scenario is to move the corresponding data slices while maintaining data consistency, transparently to the analytical tools, enabling them to use a common interface for accessing data, no matter whether this data resides in the operational data store or in the object storage.

With respect to the design of the overall architecture, the Data Store of PolicyCLOUD is conceptually a central component where data is being ingested (either via a streaming mechanism or with a static data acquisition from external sources) and is being accessed via a common interface by all analytical tools that require data retrieval for their analysis. An additional requirement is to access data that resides in external data sources that are not eligible to be physically imported to the central persistent storage of the platform remaining on premise due to data regulator constraints or due to excessive ingestion/maintenance costs. The central data store component has to provide access to such external sources, via the common interface used by the analytical tools.

At this point, it is very important to distinguish between the major three different types of data sources that the PolicyCLOUD will support: i) ingest-now data, ii) streaming data and iii) external data (external data may further be supported in future scenarios but no Use Case scenario, within the project duration, has made use of this feature). With the term stakeholder data we refer to data that belongs to the organization that can be ingested to the platform via the data acquisition mechanism. With the term streaming data we refer to data that is not static (or *data at rest*) but rather might be generated by IoT devices or coming from a social media feed such as tweeter, and requires a processing in real-time and accumulation for further analytics Finally, with the term external data we refer to both data that is either not owned by the organization and cannot be retrieved and ingested to the platform, or it cannot be ingested due to privacy considerations, and to data that might be owned by the organization, but cannot be imported due to technological constraints, and thus they are considered as external to the platform. In the following, we provide specific details on how the technology provided by the data stores and data management building block will deal with these three types of data.

- *Stakeholder data*

  In order to address the challenges for data management and overcome the barriers imposed by the data constraints coming from the use cases, the PolicyCLOUD Data Store component will rely on the LXS data repository which natively provides characteristics that are relevant to those challenges, and will be further extended in the scope of the project to cover all aforementioned requirements. More information regarding the characteristics of the datastore can found in the document of Deliverable D4.1.

- *External data*

  The challenge on the isolated silos across different kinds of data stores at each organization is addressed by leveraging the polyglot capabilities of LXS that enables to integrate its query engine with different data stores. Using the CloudMdSQL query language, which is an extension of the standard SQL, the data user can write queries in a unified manner that targets heterogeneous data stores and let the query engine of the PolicyCLOUD datastore to retrieve and merge the intermediate results. This can overcome – to the extend possible - the need for accessing data that are stored in different silos inside an organization or in external sources.

  We need to mention that no Use Case scenario required access to external data during the duration of the project. As mentioned in previous sections, PolicyCLOUD may have the ability to process external data for future scenarios after completion of the project.

- *Streaming data*

  Often it becomes necessary to manage streaming data combined with data at rest, in order to correlate events with operational data and/or update a dataset based on an event. This is a bottleneck for traditional databases when streams arrive at large scale, as they are incapable of dealing with those operational workloads at that high rate. Due to the scalable transactional processing provided by the LXS datastore and its additional interface that allows directly accessing its storage layer, it can support data ingestion coming from streams.

  Moreover, due to its extended capabilities for live aggregations, it can support the combination of streaming events with data at rest which requires data expensive operations (i.e. average value of a field) that can be supported by traditional data management systems.

Apart from the ability of PolicyCLOUD to deal with these three different types of data, the data management mechanisms of the platform will benefit from the results for the EU H2020 project BigDataStack and its Seamless Analytical Framework, where similar scenarios with regards to the movement of “cold” data from an operational to an object store are being addressed. That allows for data to be moved to the object storage at runtime, transparently to the user by ensuring data consistency and without the downgrade of the performance during the movement of the data. The data repository supports standard SQL statements via the common JDBC, and splits the data operations so that they can be executed in both underlying stores, and merges the intermediate data in order to return the same result as if the data was stored in a single database. By doing this, the data analyst does not have to alter its implementations in order to support scenarios where there is the need to combine data from both stores. In the scope of PolicyCLOUD, the prototype firstly developed in the EU BigDataStack project is being developed to cope with the scenarios defined here, which will increase its current technology readiness level (TRL).

7.5 Ethical and Legal Compliance Framework

7.5.1 Ethical and Legal Compliance Framework

To maximize societal acceptability and trust in PolicyCLOUD, and the policies developed with PolicyCLOUD’s assistance, the PolicyCLOUD consortium is aware of the need to carry out an extensive and in-depth analysis of relevant legal, regulatory, societal and ethical aspects, define appropriate requirements to address all relevant aspects identified, and pursue an optimal embedding of those requirements into the design of the solution – including a thorough evaluation to assess its success. Special attention must be paid to ethical and societal issues which may be triggered throughout the project.

Therefore, it is necessary to identify a set of controls – which may pertain, *inter alia*, to platform dimensions, features, and functionalities – and their links to the range of significant new practices enabled by the platform which are relevant from a legal, regulatory, ethical and/or societal perspective. These controls must remain aligned with the various iterations of the development of the platform, and the specificities of the different use cases which serve to demonstrate the platform’s capabilities. The **Ethical and Legal Compliance Framework** (task T3.5) thus aims at analysing and giving guidance on the legal, regulatory, ethical, and societal requirements by which PolicyCLOUD should abide.

Within these controls, particular attention is currently being paid to the choice of data sources and data extracted from those sources, as well as to the admissibility of their use by the data controllers/contributors bringing in data into the PolicyCLOUD infrastructure. This counts for “personal data”, as defined in the GDPR, but also for other types of data – such as "open data" – which may involve legal issues regarding the protection of intellectual property, including the protection of databases and trade secrets. One such control which has been put in place is a “clearing procedure” of sorts, to allow specific categories of data to be ingested by the platform based on the requirements set by end-users and the extent to which legal permission for such ingestion exists. This procedure has been implemented at the stage of "data acquisition", and is executed on data uploaded to the platform’s “interim repository”, after data is processed through the cloud gateway – after the legal/regulatory/ethical/societal assessment performed on datasets uploaded to the "interim repository", datasets which are cleared for use can be further ingested into the platform's main data repository.

Furthermore, within the controls defined in task T3.5 are controls to address the (shared) responsibilities of PolicyCLOUD, the partners and stakeholders when processing "personal data" under the GDPR, as well as basic considerations for the admissibility of use of such data. This also includes the role of the cloud provider (i.e., the organization/entity which, ultimately, will act as the provider of the PolicyCLOUD platform) as a controller or as processor, depending on the specific processing activities at stake. Under the principle of data minimization, the possibility to rely on aggregated or anonymous data (as opposed to identified or identifiable "personal data") has been explored with the relevant technical and use case partners. Moreover, guidance has been given on the requirements of security and confidentiality, as well as data protection by design and default, also considering the work performed in the context of task T3.6 (Data Governance Model, Protection and Privacy Enforcement).

Task T3.5 also addresses other ethical, as well as societal requirements, to align the platform with these to the greatest extent feasible. As such, controls have been defined to address the reliability of data and prevention of false raw data – which is of primordial importance to mitigate the risk of incorrect conclusions being derived from the platform's output, which may potentially lead to incorrect policy-making decisions. This includes considerations as to the potential for abuse of data to intentionally manipulate a decision-making process, as well as the potential for abuse of the platform as a whole. The need to ensure explainability of output provided by the platform, and that policymakers retain critical judgment when interpreting platform output and making decisions – as they retain ultimate responsibility for such decisions – is also deemed an essential ethical and societal concern, reflected in appropriate controls. Further analyses are being conducted regarding the data marketplace.

In the context of deliverable D3.3 [9][7], an analysis was carried out of the relevant legal, regulatory, ethical, and societal issues detected in relation to PolicyCLOUD, with a synthetic review of the existing debate and literature provided. These issues (the main of which were expressed above) were addressed in general terms, and from the perspective of the specific platform components and use cases, based on the information available as of the date of completion of the deliverable. From this deliverable, an initial set of controls was defined, to be used to ensure the platform's adherence to the principle of compliance by design (as better explained in the following section). This set of controls – or checklist – has been adapted over time, considering the relevant developments occurred for several components of the PolicyCLOUD platform and the Project's use cases. As it stands, this initial checklist has now been

[7] With regards to legal and regulatory issues, the scope of the analysis, in the context of this deliverable, was generally limited to EU and international law, without exploring in detail the specific national and/or local requirements related to the countries and jurisdictions in which the use cases are implemented. Nevertheless, where specific analysis on local and/or national regulations shall result as appropriate and/or necessary, this has been highlighted as a field for which further research is needed and that will be consequently developed in the next versions of the deliverable, to be released at M34.

decomposed into multiple checklists, focused on different components or aspects of relevance to the Project - as last reported in deliverable D3.6 [15], which is the most recent update to deliverable D3.3:

- The **WP2 Legal/Ethical Checklist** focuses on security, and includes a set of controls defined with reference to ENISA's EUCS – Cloud Services Scheme [17];
- The **WP3 Legal/Ethical Checklist** focuses on the platform's cloud-based infrastructure;
- The **WP4 Legal/Ethical Checklist** focuses on the platform's data repository and pre-determined analytical tools, as well as the registration process for new data sources and analytical tools;
- The **WP5 Legal/Ethical Checklist** focuses on the PDT, PME and other user-facing interfaces;
- The **WP6 Legal/Ethical Checklists** (which are divided per use case) focus on the legal/ethical requirements applicable to the different scenarios defined per use case;
- The **WP7 Legal/Ethical Checklist** focuses on the Data Marketplace.

These Checklists are used as an ongoing compliance monitoring tool within the Project. Each Checklist has been refined after various discussions with the Partners and includes a resulting set of specific controls. Each of these controls points to one or more technical and/or organizational measures which are assigned to one or more identified Partners. The Checklists are progressively updated to reflect the progress made by the relevant Partner(s) on the implementation of each measure, and what steps are left to be taken. The current status of these Checklists is last reported in deliverable D3.6 [15]; in the final iteration of deliverables D3.3 and D3.6 – which is D3.9 due on M34, a final description of all measures taken to ensure the ethical, regulatory, societal and legal soundness of the project, as defined in these Checklists, will be provided.

7.5.2 Ethical and Legal Compliance Framework Integration with Use Cases and technology

7.5.2.1 Application of the compliance by design principle

In this section, we will analyze how, during the development of the PolicyCLOUD project, compliance with the identified legal, regulatory, ethical and societal requirements is being assessed. With specific regards to data protection and privacy issues, we will also define a methodology for the implementation of a DPIA which will be conducted with regards to each of the relevant use cases.

7.5.2.2 Compliance by design approach

To address all the relevant ethical, legal, regulatory, and societal risks related to the project, a compliance by design approach is being adopted.

Compliance by design means applying a systematic approach to integrating relevant compliance requirements into tasks and processes. The effective implementation of this principle is based on the detailed and structured analysis of all the applicable requirements (as initially identified in deliverable D3.3), followed by translation of those requirements into workable compliance processes [10].

A three-stage approach is being applied:

1. The **first stage** is dedicated to the **identification and the assessment of relevant requirements**.

This was accomplished with deliverable D3.3, at the end of which an initial list of requirements/controls was defined, based on the assessment developed within the deliverable of the applicable legal, regulatory, ethical and societal issues potentially triggered by the platform's technical components and use cases.

2. The **second stage** includes the analysis **on how the rules apply to individual processes**.

This was accomplished through the breaking down of the initial list of requirements/controls into smaller "checklists" – i.e., Legal/Ethical Checklists – in which individual controls are allocated to different Work Packages within the project. As of the date of this deliverable, Checklists have been developed for the main components under the responsibility of all Work Packages (WP2, WP3, WP4, WP5 and WP7 – including also additional components which have been developed further in the meantime, such as the Data Marketplace and Incentives Management components), and for the use cases under the responsibility of WP6.

The relevant Checklists have been shared with the WP Leaders, and subsequently been refined to remove or adjust controls deemed inapplicable or unfeasible. Specific owners have been identified for each refined control, and specific, practical measures have been identified to ensure each control is implemented. Feedback has been exchanged with the WP Leaders, and the initial consolidation of the mentioned Checklists can be considered completed.

3. The **third stage** focuses on the **design and implementation of a roadmap**.

Following up on the initial consolidation of the Checklists, the owners of each specific control have been engaged to (1) describe the specific, practical measures taken (or proposed to be taken) to implement each control assigned to them, and (2) define a roadmap for the implementation of those measures, where not already implemented.

The final step in this process is to maintain continuous engagement with the WP Leaders and control owners, to assess the implementation process in accordance with the defined roadmaps, addressing any concerns which may arise over time (e.g., adjusting controls as needed to fit new platform developments).

7.5.2.3 Data protection and privacy

To address the issues related to the project and the use cases concerning data protection and privacy (i.e., for use cases where "personal data", under the GDPR, are to be processed), data protection impact assessments ("**DPIAs**", under Art. 35 GDPR) will be implemented, and the results of those assessments will be presented in the context of deliverable D3.9 of the project.

Through these DPIAs, PolicyCLOUD will assess the processing operations to be performed, as well as the technologies, tools, and systems to be used, in relation to each specific use case scenario in which the processing of personal data is envisioned, to identify inherent risks in a structured manner. Furthermore, these DPIAs will be used to identify measures which can be implemented to bring those risks down to acceptable levels. The DPIA reports will contain a systematic description of the envisaged processing operations, the purposes for which personal data will be processed, an assessment of the legitimate interests pursued (where applicable), an assessment of the necessity and proportionality of the operations in relation to those purposes, an assessment of the risks to the rights and freedoms of data subjects, and a description of the measures envisaged to address those risks, as noted in Art. 35, par. 7 GDPR [11].

These DPIAs will be performed according to the methodology defined in the international standard ISO/IEC 29134.

The process for the performance of the DPIAs will include:

1. A **preparation phase**, during which the DPIA teams will be set and provided with direction, the DPIA plan will be prepared, the necessary resources will be determined, and the relevant stakeholders will be engaged.
2. A **performance phase**, during which the information flows of personal data will be identified, the implications of the relevant use case scenario (in the context of the project) will be analyzed, the data protection and privacy risks will be assessed, a risk treatment plan will be defined and relevant privacy safeguards will be determined.
3. A **follow-up phase**, during which a DPIA report will be prepared and published, and the risk treatment plan will be implemented. In this context, also a review and/or reaudit program of the DPIA will be defined, to monitor both the correct implementation of the risk treatment plan and of the potential changes to the previously assessed personal data processing activities.

7.5.2.4 Data Security

As mentioned above, and better described in Section 2.1.1 of deliverable D3.6 [15], the Consortium has agreed that the European Union Cybersecurity Certification Scheme for Cloud Services (EUCS) [17], as a candidate certification scheme focused on the cybersecurity of cloud services, developed under the European Union (EU) Cybersecurity Act[8] by the EU's agency dedicated to achieving a high common level of cybersecurity across Europe, could serve as an appropriate standard by which to measure the security posture of the PolicyCLOUD platform and identify additional technical and/or organisational measures to be implemented.

As a result, the **WP2 Legal/Ethical Checklist** has been developed to map the EUCS requirements identified as potentially relevant to the PolicyCLOUD platform. As of the date of this deliverable, this Checklist is being further refined and discussed with the relevant Partners so as to determine whether any identified requirements are irrelevant or unfeasible, or whether any additional EUCS requirements should be

[8] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013, available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32019R0881.

considered), and will subsequently be used to track and document the implementation of the consolidated requirements by the relevant Partners (identified as owners for each requirement in discussions with the Consortium) until the end of the PolicyCLOUD project.

7.5.2.5 Data Subject Rights

As better described in Sections 2.2.4, 2.3.4, 2.2.5 and 2.3.5 of deliverable D3.6 [15], the **WP3 Legal/Ethical Checklist** and **WP4 Legal/Ethical Checklist** have been expanded to include a description of the rights afforded to individuals under the GDPR, and an indication of the technical abilities which the PolicyCLOUD platform should allow (either under individuals' autonomous control, or under the control of platform system administrators) to ensure that these rights can be appropriately exercised.

It has been confirmed by the relevant Partners that these abilities can be covered via the PolicyCLOUD platform, at least through manual intervention by system administrators, regarding (1) personal data held on PolicyCLOUD platform users, and (2) personal data held on individuals within data sources processed via the PolicyCLOUD platform. This includes the possibility to implement a mechanism for deletion of such personal data within the PolicyCLOUD platform, as it appears that the cloud-based infrastructure on which the platform is hosted does not present any relevant technical obstacles to the implementation of this ability (which is relevant also for enforcing data retention periods which may ultimately be defined for retained personal data after the conclusion of the Project).

7.5.2.6 Analytics Compliance

As better described in Section 2.3.1 of deliverable D3.6 [15] and Section 2.3.2 of deliverable D4.3 [18], to assist in ensuring appropriate auditability of all analytics tools leveraged on the PolicyCLOUD platform, a standard logging service has been implemented in the platform, as a centralised, PolicyCLOUD project-wide component. This service also addresses security concerns, by allowing, *inter alia*, for the identification of security threats, analysis of suspected incidents (i.e., forensic analysis), monitoring of internal policy violations, collection of information on abnormal events and debugging both performance and functional issues. This service will, nonetheless, be further assessed as part of the project's efforts to harmonise the approach to data security around relevant standards – in particular, the EUCS (mentioned above). The service has been designed specifically regarding the pre-existing analytics tools implemented within the PolicyCLOUD platform, but can also potentially be extended to additional analytics tools which may be registered for use via the platform.

As reported in Section 2.1.2 of deliverable D4.3 [18], regarding both pre-existing and additional analytics tools, as a practical means to provide assurances as to their legal/ethical soundness for PolicyCLOUD users, **specific fields/parameters have been added to the registration Application Programming Interfaces (APIs) on the platform**, which are to be populated with the following details regarding each individual analytics tool:

- *biasDoc*. This parameter permits, and in fact could even oblige (e.g., by checking that this parameter has a minimal required length and/or that an attachment of a defined format and file size range is included), registrants to link bias management information/documentation to the tool upon registration. This should provide details on specific measures taken to address the risk of biases inherent to the functioning of the tool. It could also permit / oblige registrants to upload or indicate one or more test datasets on which the tool can be applied, to demonstrate that resulting bias measurements do not exceed a given maximum (in other words, that the specific measures taken to address the risk of biases are functioning effectively, as intended);

- *tradeoffsDoc*. This parameter permits, and in fact could even oblige (e.g., by checking that this parameter has a minimal required length and/or that an attachment of a defined format and file size range is included), registrants to link trade-off management information/documentation to the tool upon registration. This should provide details on the relevant trade-offs encountered in the development of the tool, decisions made concerning the balancing of competing requirements (e.g., result precision vs. fairness) and measures taken to implement and document those decisions.

As of the date of this deliverable, the specific Partners responsible for each relevant tool have been engaged to provide information about the training and testing protocols/programs, mechanisms facilitating auditability of AI-based systems (including the traceability of the development process, the sourcing of training data used, the logging of processes/outcomes/impacts), the trade-offs between applicable legal/ethical requirements and principles considered during design, and the degree of possibility of false positive/negative correlations, applicable (or to be applied, where relevant) to the tools for which they are responsible. This was done through the development of a questionnaire shared with such Partners and subsequent discussions on the answers provided. The final outcome of this exercise will be used to provide recommendations to the respective Consortium members so as to maximise the assurances provided as to the legal/ethical compliance of such tools, as well as to populate the abovementioned fields/parameters on the PolicyCLOUD platform, and will be reported in deliverable D3.9.

7.5.2.7 Data Source Registration

As better described in Section 2.3.3 of deliverable D3.6 [15] and Sections 2.3.3 of deliverable D4.3 [18], a similar solution to that found for analytics tools has been implemented for datasets/data sources registered (or to be registered) on the PolicyCLOUD platform, in that specific fields/parameters have been added to the registration Application Programming Interfaces (APIs) on the platform, which are to be populated with the following details regarding each individual data source:

- *biasDoc*. This parameter permits, and in fact could even oblige (e.g., by checking that this parameter has a minimal required length and/or that an attachment of a defined format and file size range is included), registrants to link bias management information/documentation to the data source upon registration. This should provide details on the bias detection methods applied to the data source, the specific biases identified, and the specific measures taken to address any such biases.
- *GDPRDoc*. This parameter permits, and in fact could even oblige (e.g., by checking that this parameter has a minimal required length and/or that an attachment of a defined format and file size range is included), registrants to link privacy / data protection management information/documentation to the data source upon registration. This should provide details on whether any personal data is included within the data source and, if so, the measures taken to

ensure that the data source can be leveraged via the PolicyCLOUD platform in compliance with the EU privacy/data protection legal framework, as defined by the GDPR and other applicable data protection laws (e.g., to ensure compliance with the principles of lawfulness, transparency and purpose limitation under the GDPR, considering the purposes for which the personal data within the data source were originally collected and the purposes for which those personal data will be further processed via the platform).

- *authDoc*. This parameter, permits, and in fact could even oblige (e.g., by checking that this parameter has a minimal required length and/or that an attachment of a defined format and file size range is included), registrants to link information/documentation to confirm and/or demonstrate that the registration of the data source has been authorised by relevant rightsholders (or that such authorisation is not required under the EU legal framework).

Though initial legal/ethical soundness assessments have been carried out on the different data sources which have been uploaded for further processing onto the PolicyCLOUD platform so far – see Section 2.5 of deliverable D3.6 [15] for more on this - the different use case Partners (in charge of identifying the relevant data sources for the PolicyCLOUD project) will be further engaged to provide information on each identified data source, in order to populate the abovementioned fields/parameters on the PolicyCLOUD platform. This information will be further reported in deliverable D3.9.

7.5.2.8 PDT, PME and Data Marketplace: Privacy Policy and Terms and Conditions

As better described in Section 2.4.1, 3.1.2.2 and 3.2.1 of deliverable D3.6 [15], it is important to define terms and conditions for the use of the user-facing tools included with the PolicyCLOUD platform – the PDT, PME, Incentives Management system and the Data Marketplace – so as to properly regulate the service relationship established with PolicyCLOUD platform users (i.e., individual users, or organisations to which the individual users belong). These terms and conditions would need to be accepted for the use of the PDT/PME/Data Marketplace to be allowed.

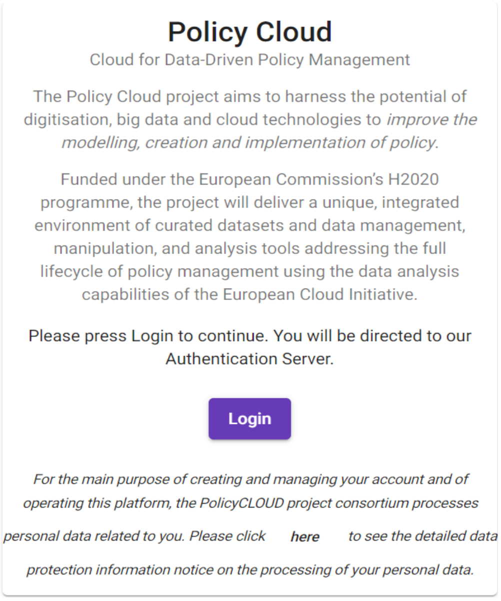
As of the date of this deliverable, a first version of terms and conditions for the PDT (covering the PME and Incentives Management system) and the Data Marketplace has been developed, and is currently under discussion with the relevant Partners for implementation on the respective tools.

In parallel, a privacy policy – referred to as the End Users Data Protection Information Notice – has been implemented on the PDT/PME. The goal of this policy is to provide written information to PolicyCLOUD platform users as to how their personal data may be handled when using the PDT/PME in a concise, transparent, intelligible, and easily accessible form, using clear and plain language, meeting all the requirements of Arts. 13 and 14 GDPR. More specifically, the data protection information notice is provided to PDT users using a two-layer approach:

- The first layer is represented by a pop-up banner which is shown to users when accessing the PDT. This pop-up banner includes a link to the second layer data protection information notice (i.e., the extended version of the data protection information notice). The text of the pop-up banner is provided in D8.1. See also Figure 6 for more details.
- The second layer is represented by the extended version of the data protection information notice, including all elements required by Arts 13 and 14 GDPR. This extended version of the data

protection information notice, the text of which is provided under D8.1 [19], is accessible to PDT users by clicking on a footer named "End Users Data Protection Information Notice" published on all the pages of the PDT's web environment. See also Figure 7 and Figure 8 for more details.

Additionally, it is important to note that the PDT and PME do not use cookies or similar tracking technologies, so no specific compliance requirements are needed to this regard.



Policy Cloud

Cloud for Data-Driven Policy Management

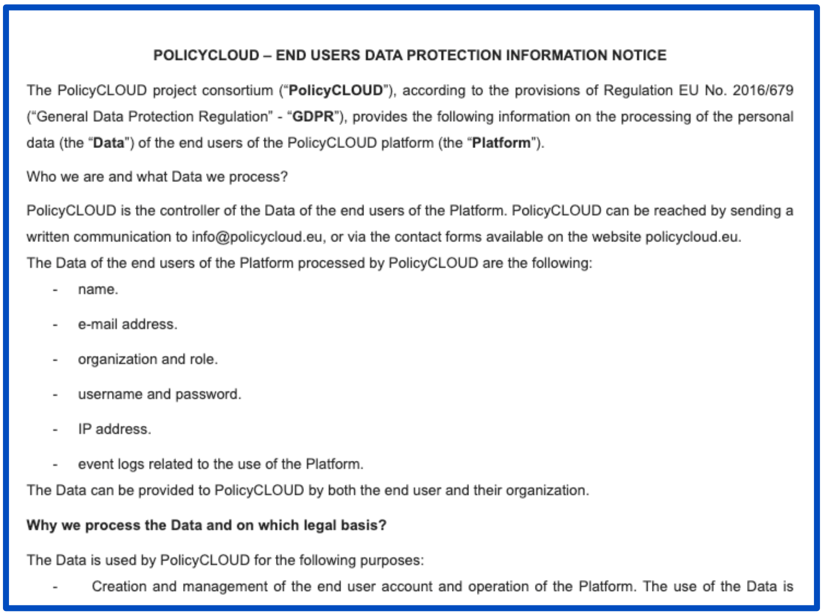
The Policy Cloud project aims to harness the potential of digitisation, big data and cloud technologies to *improve the modelling, creation and implementation of policy.*

Funded under the European Commission's H2020 programme, the project will deliver a unique, integrated environment of curated datasets and data management, manipulation, and analysis tools addressing the full lifecycle of policy management using the data analysis capabilities of the European Cloud Initiative.

Please press Login to continue. You will be directed to our Authentication Server.

For the main purpose of creating and managing your account and of operating this platform, the PolicyCLOUD project consortium processes personal data related to you. Please click ***here*** *to see the detailed data protection information notice on the processing of your personal data.*

FIGURE 6 – POP-UP BANNER



POLICYCLOUD – END USERS DATA PROTECTION INFORMATION NOTICE

The PolicyCLOUD project consortium ("**PolicyCLOUD**"), according to the provisions of Regulation EU No. 2016/679 ("General Data Protection Regulation" - "**GDPR**"), provides the following information on the processing of the personal data (the "**Data**") of the end users of the PolicyCLOUD platform (the "**Platform**").

Who we are and what Data we process?

PolicyCLOUD is the controller of the Data of the end users of the Platform. PolicyCLOUD can be reached by sending a written communication to info@policycloud.eu, or via the contact forms available on the website policycloud.eu.
The Data of the end users of the Platform processed by PolicyCLOUD are the following:

- name.
- e-mail address.
- organization and role.
- username and password.
- IP address.
- event logs related to the use of the Platform.

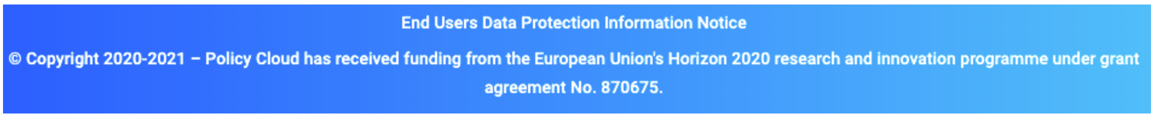
The Data can be provided to PolicyCLOUD by both the end user and their organization.

Why we process the Data and on which legal basis?

The Data is used by PolicyCLOUD for the following purposes:

- Creation and management of the end user account and operation of the Platform. The use of the Data is

FIGURE 7 – END USERS DATA PROTECTION INFORMATION NOTICE


End Users Data Protection Information Notice

© Copyright 2020-2021 – Policy Cloud has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870675.

FIGURE 8 – END USERS DATA PROTECTION INFORMATION NOTICE (FOOTER)

7.6 Layer 2 - Data Acquisition and Analytics

Components: Data Cleaning (T4.2), Data Interoperability (T4.2), Data Fusion (T4.1), Situational Knowledge Analysis (T4.3), Opinion Mining (T4.4), Sentiment Analysis (T4.4), Social Dynamics (T4.4), Behavioral Analysis (T4.5), Optimization and Reusability (T4.6)

7.6.1 Data Acquisition and Analytics – Positioning & Goals

In this section we provide the high-level architecture of the Data Acquisition and Data Analytics tasks, which is responsible for ingesting the data from various sources while applying filtering and initial analytics, and preparing it for deeper analytics on longer term storage (DB, object storage). The Requirements for Reusable Models and Analytical Tools are provided in Deliverable D2.5 [23] (section 5.3).

The relevant part from overall architecture is shown in Figure 9 for convenience. This part focuses on Data Acquisition and Data Analytics over which the integrated processing will be applied.

More specifically, data fusion tasks are integrated with the initial analytics and data processing tasks (e.g. filtering, validation and cleaning). Applying deeper analytic tasks are performed in collaboration with the continued data fusion (e.g. moving older data from DB to object storage).

From the aspect of work packages partitioning, this layer is under the responsibility of WP4 (Reusable Models & Analytical Tools) and its tasks, with a strong relation to Task 3.3 (Cloud Gateways) and Task 3.6 (Data Governance Model, Protection and Privacy Enforcement) of WP3. In Figure 10 we show the conceptual model from the work packages partitioning point of view and the WP4 interfaces to WP3 below and WP5 above, as provided in the Grant Agreement document.

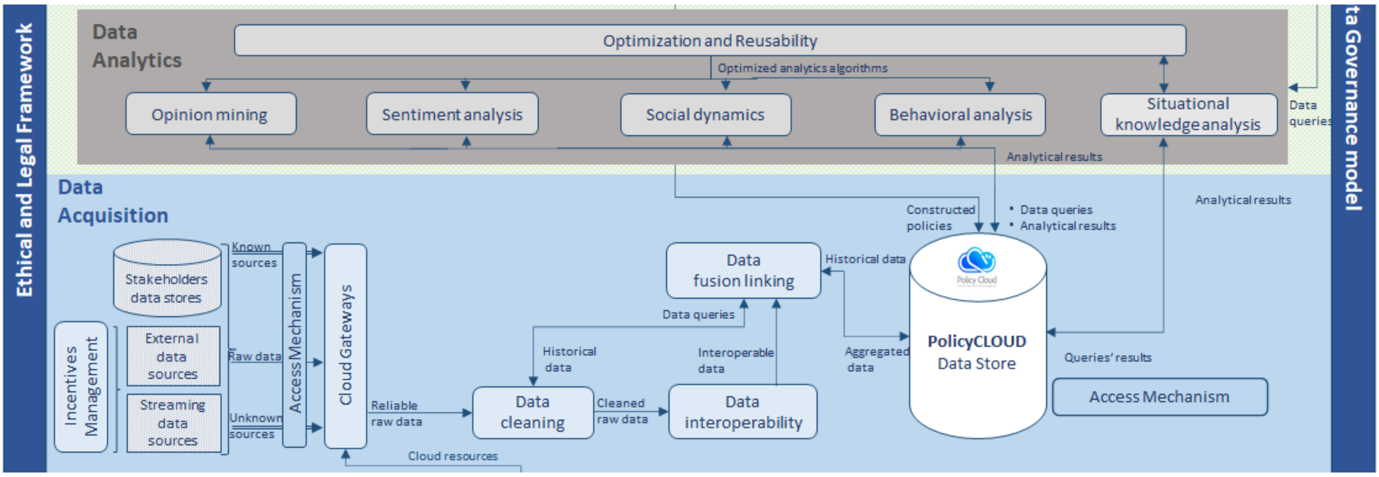


FIGURE 9 – PART OF THE POLICYCLOUD OVERALL ARCHITECTURE DIAGRAM RELEVANT TO WP4

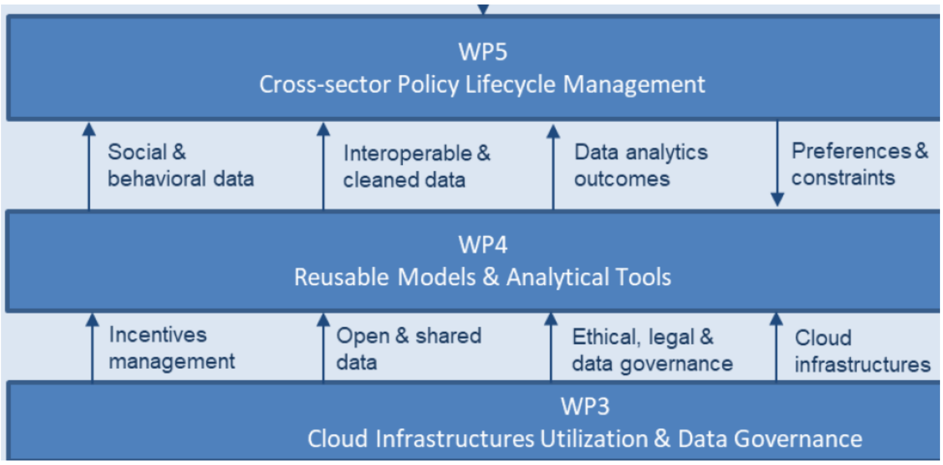

WP5
Cross-sector Policy Lifecycle Management
Social & behavioral data
Interoperable & cleaned data
Data analytics outcomes
Preferences & constraints
WP4
Reusable Models & Analytical Tools
Incentives management
Open & shared data
Ethical, legal & data governance
Cloud infrastructures
WP3
Cloud Infrastructures Utilization & Data Governance

FIGURE 10 – WP4 INTERFACE WITH WP3 AND WP5

The major goals of Data Acquisition and Analytics layers are on par with WP4 defined goals:

- Data fusion and aggregation – for different data sources types.
- Data cleaning ensuring quality of information, sources reliability assessment, reliability-based selection of information sources.
- Sentiment analysis techniques for policy assessment.
- Analysis of the social and behavioural data and requirements provided by social science experts for data selection in a given case.
- Decoupling of the analytical models and tools from the underlying infrastructure and datastores, assuring their reusability.
- Enabling the architecture to support the plugin of additional analytical tools.

7.6.2 Extensibility and Reusability of Analytic Functions

The architecture of the Data Acquisition and Analytics layers will provide extensibility and reusability of analytic functions. New analytics functions (services) can be registered into PolicyCLOUD and reused for applying analytics on new and existing registered data sources. The decided alternative at this point is a registration as serverless functions that are activated on demand, either by a direct PolicyCLOUD user request or by event/rule. There are two types of functions:

1. Ingest analytics / transformation function, which will be used to apply initial analytic and/or transformation on the data fusion path of data sources.
2. Analytic function which can be applied to data at rest which will be activated upon PolicyCLOUD user action on specified data source (which was already ingested) to provide analytic results for policy decisions.

The design details of analytic functions registration and activation are provided in deliverable D4.1.

7.6.3 Data Cleaning

The Data Cleaning component offers all the appropriate algorithms and techniques for detecting and correcting (or removing) to the maximum extend possible, corrupt or inaccurate records from the collected data that are retrieved as an input from the Cloud gateways component. More specifically, this component is responsible for identifying all the incomplete, incorrect, inaccurate or irrelevant parts of this data, and then replacing, modifying, or deleting the dirty or coarse data. Thus, possible missing, irregular, unnecessary, or inconsistent data are found and totally cleaned. Especially dealing with missing data is one of the most tricky but common parts of the data cleaning process since most of the models do not accept missing data. To this context, the Data Cleaning component detects and totally cleans all the missing data by combining techniques such as the Missing Data Heatmap, the Missing Data Percentage List, as well as the Missing Data Histogram, thus extracting quite accurate and reliable results. With regards to irregular data, cleaning is made possible by using techniques such as the Histogram and the Descriptive Statistics for the numeric values, and by exploiting the Bar Chart for categorical values.

Regarding the unnecessary data, since it refers to data that will not add any value to the PolicyCLOUD overall platform, by constructing the corresponding rules and constraints, all the uninformative/repetitive, irrelevant values, as well as the duplicates are automatically detected and may be erased. Finally, since any possible inconsistent data are automatically corrected it is also crucial that all the collected datasets will follow specific standards to fit the corresponding PolicyCLOUD data models. As soon as all the data is fully cleaned they are sent into the Data Interoperability component for further utilization.

7.6.4 Data Interoperability

The Data Interoperability component aims to enhance the interoperability of analytics processing in the PolicyCLOUD project based on data-driven design, coupled with linked data technologies, such as JSON-LD [12], and standards-based ontologies and vocabularies to improve both semantic and syntactic data and dataset interoperability. The provided Interoperability Component seeks to extract semantic knowledge and good quality information from the cleaned data that will be the input to its system, as shown in the initial architecture of the overall project. This knowledge, shaped in a machine-readable way, will be used in next tasks for Big Data analytics, Opinion Mining, Sentiment Analysis etc.

One of the preliminary steps of this component is to identify relevant, publicly available, and widely used classifications and vocabularies, such as the Core Person Vocabulary provided by DCAT Application Profile for Data Portals in Europe (DCAT-AP), that can be re-used to codify and populate the content of dimensions, attributes, and measures in the given datasets. Hence, this component aims to adopt standard vocabularies and classifications early on, starting at the design phase of any new data collection, processing or analytical components. Using for example NLP techniques and tools like Text Classification, NER, POS tagging and even Machine Translation [13] [14] we can identify and classify same entities, their metadata and relationships from different datasets and sources and finally create cross-domain vocabularies in order to identify every new incoming entity. Likewise, in order to create and enhance semantic interoperability between classifications and vocabularies this component seeks to engage in structural and semantic harmonization efforts, mapping cross-domain terminology used to designate measures and dimensions to commonly used, standard vocabularies and taxonomies. Thus, by implementing a "JSON-LD context" to add semantic annotations to interoperability component's output, the system will be able to automatically integrate data from different sources by replacing the context-depended keys in the JSON output with URIs pointing to semantic vocabularies, that will be used to represent and link the data. This mechanism enhances information by connecting data piece by piece and link by link, allowing for any resource (authors, books, publishers, places, people, hotels, goods, articles, search queries) to be identified, disambiguated and meaningfully interlinked.

7.6.5 Data Fusion with Processing and Initial Analytics

While the PolicyCLOUD environment has been architected to support data fusion as shown in Figure 11, we have not proceeded into its implementation. The reason for this, is that existing scenarios and policy makers' requirements gave more emphasis and prioritized other aspects of the system such as integration with external tools. Taking into consideration that data fusion is supported by the existing architecture, its implementation will be considered during the exploitation phase of PolicyCLOUD, after completion of the project, based on additional scenarios/use cases for which data fusion will be more suitable.

In the following paragraphs we provide the architecture for integration of all the tasks relevant to data fusion. We demonstrate this integration by an end-to-end example data fusion scenario, from a Twitter social network data source. The data is fused, cleaned, validated and initially analysed for extracting the relevant knowledge insights which are then persistently stored for future deeper analytics and possibly generating immediate alerts. The participating tasks in this scenario are:

- T3.3 Cloud Gateways.
- T4.1 Cross-sector Data Fusion Linking.
- T4.2 Enhanced Interoperability & Data Cleaning.

- Potential initial analytics by T4.3 Situational Knowledge Acquisition & Analysis, T4.4 Opinion Mining & Sentiment Analysis and T4.5 Social Dynamics & Behavioral Data Analytics.

The framework for data fusion and analytics will either be based on Apache Spark Streaming open source[9] , KSQL[10] , or Serverless engine based on Apache OpenWhisk[11]. In Figure 11 we depict the end-to-end data path for this scenario.

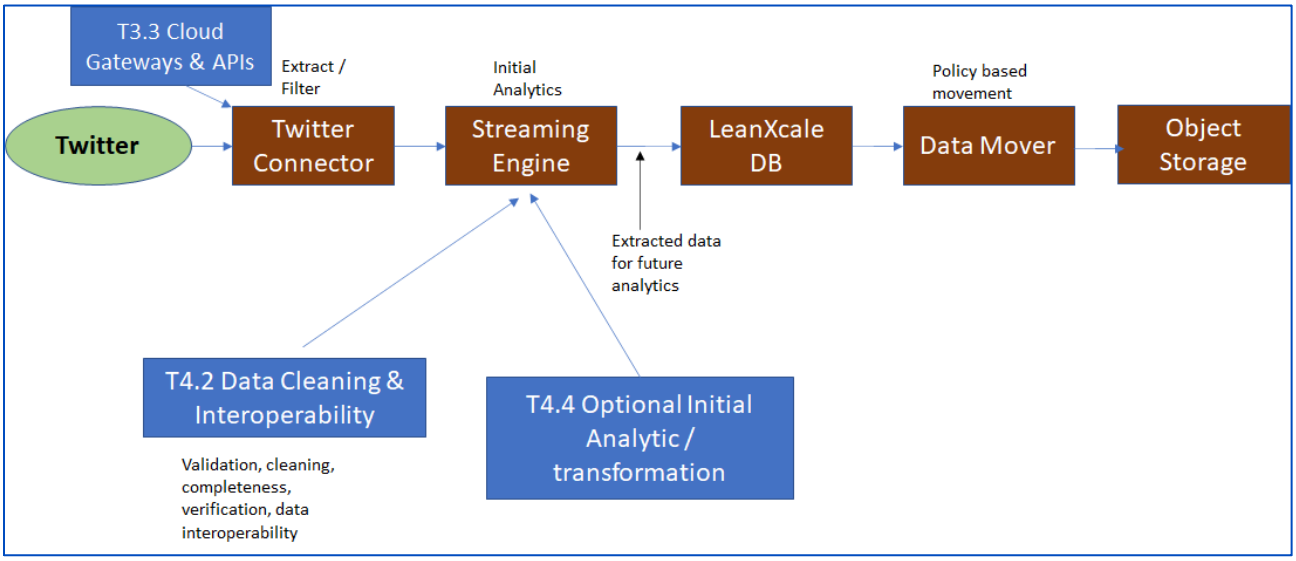

T3.3 Cloud Gateways & APIs
Extract / Filter
Initial Analytics
Policy based movement
Twitter
Twitter Connector
Streaming Engine
LeanXcale DB
Data Mover
Object Storage
Extracted data for future analytics
T4.2 Data Cleaning & Interoperability
Validation, cleaning, completeness, verification, data interoperability
T4.4 Optional Initial Analytic / transformation

FIGURE 11 – THE STREAMING DATA PATH

Task 4.1 (Cross-sector Data Fusion Linking) is responsible for the overall data path and streaming framework in this scenario. The Twitter connector has been implemented by task T3.3 (Cloud Gateways) and creates the stream of relevant data into the Streaming engine. It is expected to apply basic filtering by policy rules that are active in the PolicyCLOUD framework (resulting from actual policies that are subject for validation). The data cleaning and reliability validation was performed by Task T4.2 which provided analytic tools that are run within the streaming pipeline. Optional initial analytics on the streamed data may be performed by tasks T4.4 (Opinion Mining & Sentiment Analysis).

At the end of the data path, the Data Mover is responsible for moving historic data slices from hotter storage (DB) to a colder (object storage). This is performed periodically, according to certain policy rules (discussed more in details in the next section).

7.6.6 Seamless Analytics on Hybrid Data at Rest

In this section we provide the architecture for applying the analytics functions on the data at rest, which is combined of knowledge insights extracted within the data fusion, as well as more 'raw' data (however

[9] https://spark.apache.org/streaming
[10] https://github.com/confluentinc/ksql
[11] https://openwhisk.apache.org

still after cleaning and validation processes). The "right" side of the data path in Figure 12 present a periodical movement of older data from hotter storage (DB) to a colder (object storage) according to policy rules, which address the scalability and cost aspects of dealing with big data. Object storage is the perfect platform for storing big data for analytic purposes when no future modification of the data is expected, while the DB platform is superior performance-wise for analytics on the hotter data. The requirement is to apply seamlessly analytics on both hot (in the DB) and cold (in the object storage) data. The basic technology of data movement and seamless analytics was developed by IBM and LeanXcale partners in the BigDataStack H2020 project[12] and is exploited and adapted for PolicyCLOUD.

The participating tasks for the provided functionality are:

- T4.1 Cross-sector Data Fusion Linking
- T4.3 Situational Knowledge Acquisition & Analysis
- T4.4 Opinion Mining & Sentiment Analysis
- T4.5 Social Dynamics & Behavioral Data Analytics
- T4.6 Optimization & Reusability of Analytical Tools

As depicted in Figure 12 the framework for data movement and seamless analytics will be provided by overall task T4.1 (Cross-sector Data Fusion Linking). Task T4.6 (Optimization & Reusability of Analytical Tools) Optimization aspects (to be developed in the later phases of the project) will additionally provide the interface for seamlessly applying the analytic tasks as T4.4 (Opinion Mining & Sentiment Analysis) and T4.5 (Social Dynamics & Behavioral Data Analytics) on the data at rest.

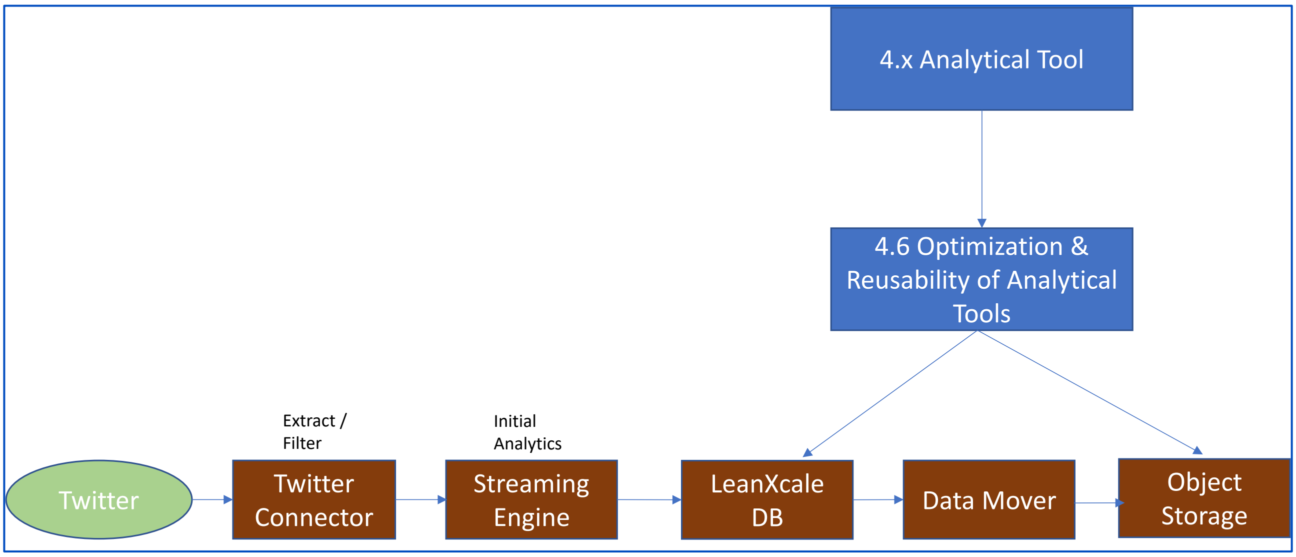


FIGURE 12 – SEAMLESS ANALYTICS ON INGESTED DATA

[12] https://bigdatastack.eu

7.6.7 Situational Knowledge Analysis

In the context of PolicyCLOUD the Situational Knowledge Acquisition (SKA) component brings the capability of acquiring knowledge from the Data & Policy aspects of the platform. The extracted knowledge is used to influence the decisions taking place based on the PolicyCLOUD system.

The following capabilities will be provided through the SKA component:

- It deals with real-time facts (such as data from sensors) from which it derives situational knowledge.
- A situational knowledge model (SKM) be provided for structuring the knowledge acquired. This data model contains a high-level description of real-word situations (context) which are the interest of the PolicyCLOUD system. The model is defined by the use cases based on the types of situations/context to be acquired.

Some of the characteristics of the component are:

- Feature Extraction. The extraction knowledge stage is done through Feature Extraction (ML) techniques able to create/derive new situational features from existing ones. This extraction step enhanced by the situational knowledge model which guide the derivation of new features or the abstraction of existing ones.
- Dataset clustering and categorization. Data categorization must be possible in a very flexible way according to the structure envisaged for formal descriptions of business fitted entities [15] (Olszewski, Robert, 2001).

7.6.8 Opinion Mining

The following tasks have been identified as being the basic activities to be performed in the context of opinion mining and sentiment analysis. The identification of these tasks is the result of internal conversations with use case owners, in order to extract information and needs for data analytics based on the various scenarios.

- *Opinion Mining.* Observe events and social attitude in respect to specific topics.
- *Named-entities recognition.* Identification of specific entities (users, locations, groups, …) cited on text.
- *Graph Analysis.* This task will develop an additional component that will perform further analytics by generating a “contributor graph” based on the contributors that are talking about the policies. This graph can be built on top of any platform with enough information about the contributors (e.g. Twitter), to determine the main influencers and create groups of similar contributors. This requirement will be refined based on the data that will be provided by each pilot. Other mechanisms such as page-rank, will be developed to generate the common analysis on graphs.

A specific focus will be devoted to particularities of social networks, such as:

- *Hashtags Detection,* identification of Twitter style hashtags from text.

- *Twitter Hashtags and Mentions Tacking,* find and monitor mentions on Twitter regarding specifics hashtags or topics.
- *User Monitoring,* identification and monitoring of most popular users who comment about specific hashtags or topics.

Additional analysis such as social media-based Location Surveillance or Topic-related expressions identification (identification of new words or expression which might have hidden relationships with known ones) can be also objective of T4.4 task.

This component will follow the same approach as the sentiment analysis component using Apache NiFi to create a pipeline in a modular way to achieve the described objectives.

7.6.9 Sentiment Analysis

This component performs a sentiment analysis based on the input received from the pilots about their policies. This input comes from what the citizens say in social media channels, from platforms owned by the pilot (getting feedback on various subjects), or other channels that have been discussed through the duration of the project. Having this input as also additional information extracted about a specific topic (such as which entities are involved), a sentiment is assigned (Positive, Negative, or Neutral). To achieve this, it is needed to train the sentiment models with different types of data from different scenarios in order to receive the best accuracy possible.

The development of this component takes advantage of powerful tools such as Apache NiFi, in order to create pipelines in an easy and modular way to be adapted to vary situations without the necessity of repetitive working. It has a common NLP part to analyse the text arriving as an input from different sources (social media, text files, or others). The sentiment value assignment for each text is stored in the database provided by PolicyCLOUD to be used by other components.

7.6.10 Social Dynamics

The Social Dynamics component consists of a concurrent, web-based environment for social simulation. The environment allows the user to create graph-based population models online. These models satisfy various parameters set by the user in terms of size, individual characteristics affecting social behavior, link characteristics, individual and connection dynamics. In addition, it is feasible to upload appropriately structured population data from databases conforming with these parameters. Individual characteristics consist of sets of variables that capture the relevant attributes for each individual in the model. Link characteristics specify a set of variables used for the creation of weighted links between individuals. Individual dynamics consist of a set of rules describing the conditions under which individual characteristics can change and the ways these changes can affect individual characteristics. In an analogous way, connection dynamics consist of a set of rules describing the conditions under which link weights can change and the ways these changes can affect link characteristics. A special-purpose modelling language will be developed that allows users to specify all these parameters online in the simulation environment. Based on these specifications, the environment is able to simulate in real-time the dynamics of such populations and store the results in a database for further processing by interested parties. The environment exploits opportunities for the breakdown of the tasks in each simulation into concurrent units that allow the simulator to optimize its use of computational resources. Integration of the developed Social Dynamics framework with PolicyCLOUD is presented in sections 7.6.11.3 and 7.6.11.4.

7.6.11 Data Acquisition and Analytics Integration

7.6.11.1 Architecture Integration

The containerized environment has been developed at https://indigo-paas.cloud.ba.infn.it .

The following components have been installed as containers on the environment:

- OpenWhisk
- Kafka
- Leanxcale datastore
- A Cloud Gateway

An initial GitLab instance at https://registry.grid.ece.ntua.gr/ was used (before it was also moved to the cloud environment) providing repositories for the project code and containers.

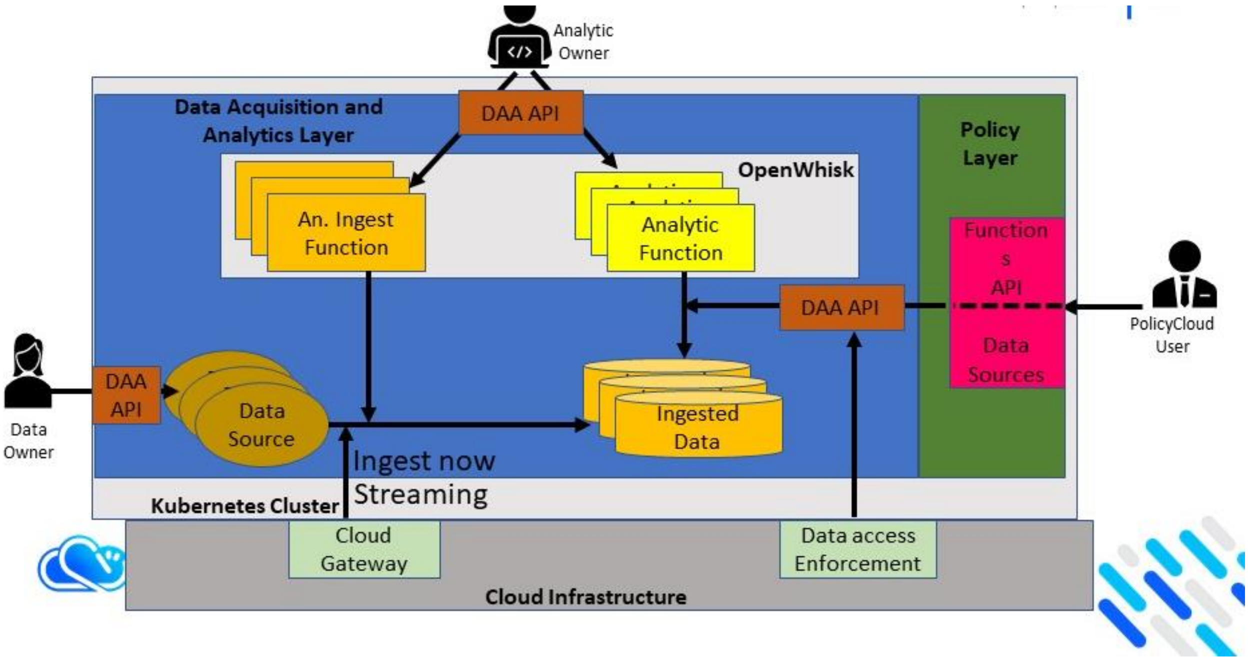


FIGURE 13 - WP4 CONSTITUENT ARCHITECTURE

The architecture implemented shown in Figure 13 demonstrates the following capabilities:

1. Registration of analytic ingestion functions which are implemented as OpenWhisk serverless functions. The ingest functions can be written in any OpenWhisk supported programming languages (e.g. Java, Python, node.js).

2. Registration of data sources (stream/ingest-now). From these data sources, data is imported to PolicyCLOUD's backend datastore (the LXS database) while being transformed by the ingest functions described in the previous paragraph. In the process, Kafka is used for buffering.

3. Analytic functions can be registered and then run on ingested data. The data is read from the data store (the LXS database), then processed while the output is presented to the PolicyCLOUD user who has initiated the function invocation.

7.6.11.2 INTEGRATION WITH THE KUBERNETES CLUSTER

Integration of OpenWhisk, Kafka and LeanXcale database with Kubernetes has been successfully achieved based on extensive tests during. This integration has been fundamental for the successful implementation of the PolicyCLOUD environment.

7.6.11.3 INTEGRATION OF THE SOCIAL DYNAMICS COMPONENT

The integration of the social dynamics component with the overall PolicyCLOUD environment has been concluded. Social dynamics was running as a stand-alone web-based component. This component is an interactive meta-simulation tool as it provides a modelling environment for developing social simulations, it automatically executes simulations and it reasons about the suitability of different policy alternatives. To this end, it accepts/generates a variety of inputs (policy models, high volumes of population data, simulation models, evaluation criteria for policy alternatives), it generates high data volumes as outputs, it requires significant computational resources (memory/cores) for simulation execution, and, because of its interactive nature, it assumes not a one-shot but a constant interaction with its user to develop, run and evaluate alternatives. We have performed an analysis examining the pros and cons of integrating such a component in a serverless paradigm versus using a virtual server for this component from a cloud provider that can communicate via messages with the data store and the rest of the analytics components. Our analysis has been performed in terms of the resource, scheduling and reliability requirements, along with the response times that will be feasible under a constant user interaction scenario in both integration alternatives. As a result, Politika has been integrated as an external framework as described in sections 7.6.11.4 and 7.6.11.5 that follow.

7.6.11.4 INTEGRATION WITH EXTERNAL FRAMEWORKS

For various reasons it may not be feasible to fully integrate external frameworks within the PolicyCloud framework. For instance, too much effort may be required to port an existing framework to the serverless paradigm. It also happens in some cases that the serverless paradigm is not well suited to the external framework. Despite these challenges, we have made it possible to integrate such external frameworks with PolicyCLOUD. The full description of this integration can be found in deliverable D4.5 section 2.2.5 while in this document we present an outline focusing on how such an external framework integrates with the overall architecture.

The Politika tool, performing Social Dynamics analysis, is an example of such an external analytical framework and is used as a paradigm for the integration of external frameworks with PolicyCLOUD. In order to facilitate the integration of Politika with PolicyCLOUD and with any external framework, we have developed a generic function that acts as the bridge between the PolicyCLOUD platform and the external analytical framework. This way, the PolicyCLOUD platform can interact with this function in the same manner as with any other native function.

The only requirement for the external framework is to implement a REST API which exposes the functionalities required by the end-users of the PolicyCLOUD. The REST web methods are invoked by the generic bridge function in order to allow the interaction between the two platforms. The generic bridge function invokes the REST web method by providing a list of values of the input parameters that have been previously defined by the provider of the external analytical framework. In the case of Politika, after the tool receives these input parameters, it feeds them to run the simulation, and upon completion it returns a list of output values that will be stored in the PDT of the PolicyCLOUD, by the generic bridge function. This allows for the result of the Politika simulation to be visualized using the PDT graph dashboards.

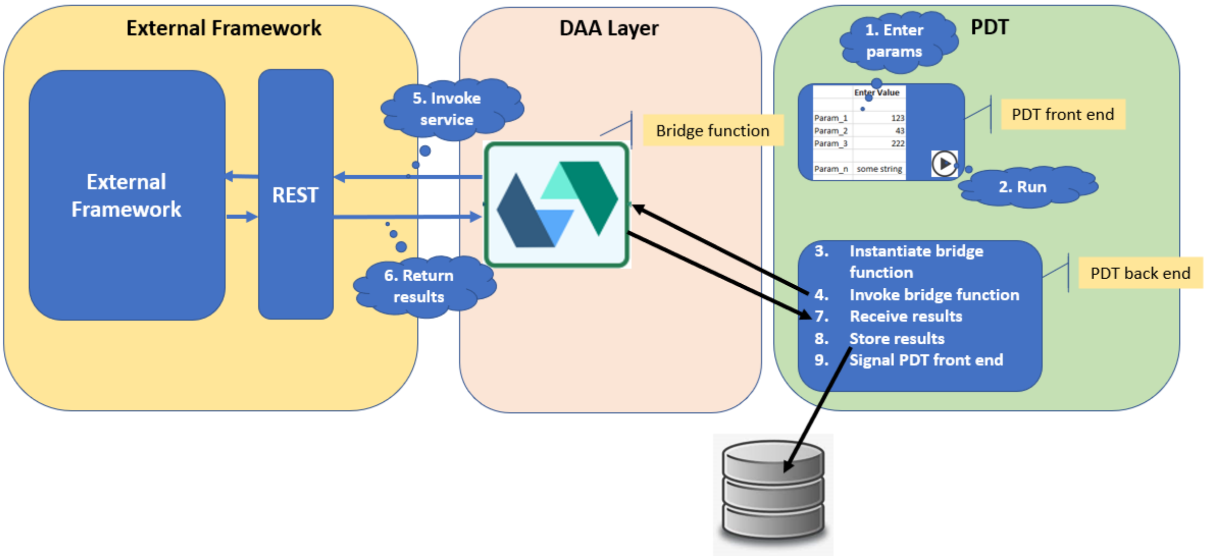


FIGURE 14 - INTEGRATING ARCHITECTURE FOR EXTERNAL FRAMEWORKS

In the following list, we detail the various stages of the invocation of an external framework from PolicyCloud, depicted also in Figure 14:

1. As a first step, the policy maker enters at the PDT the input, as defined by the external framework provider during the registration phase.
2. The policy maker invokes the corresponding function by pressing a "run" button.
3. The PDT front-end informs the backend that an invocation of a specific function needs to take place with a given user input.
4. The backend constructs the JSON body with all parameters and interacts with the DAA layer to deploy and execute the bridge function into the Openwhisk serverless platform passing it the entered parameters.

5. The DAA executes (through OpenWhisk) the requested function, providing it the input parameters given by the end-user, along with all its related configuration arguments that had been defined during the registration process.
6. In the case of an external analytical framework, the bridge function retrieves the URL endpoint of the REST interface of the framework, and synchronously invokes this REST entry point passing it all the parameters defined by the end-user during the runtime.
7. The external framework upon invocation runs the specified service using the passed parameters.
8. The external framework sends back to the invoking bridge function the results of the service invocation(s).
9. The bridge function receives the answers and stores them within the database, using the PDT REST API, in a similar manner as for the regular DAA analytic functions.
10. After the PDT backend persistently stores the output results, it notifies the frontend via messages sent through websockets.
11. The PDT front end after receiving the signal invokes the visualization function corresponding to the received results which the policy maker can now analyze.

In summary, the above steps are identical with the sequence flow of the invocation and execution of a function from the PDT to the serverless platform, as it has been described in the WP5 related deliverables. The only difference is that our novel bridge function, instead of being executed locally, consuming data from the PolicyCLOUD datastore, communicates with the external analytical framework via its REST endpoints, delegating to it the execution of the desired function and finally retrieves the results in the body of the HTTP response of the REST API.

7.7 Layer 3 – Policy Management Framework

Components: Policies Modelling (T5.2), Policies Implementation (T5.1), Policies Clusters (T5.4), Policies Experimentation (T5.5), Policies Evaluation (T5.6)

7.7.1 Policy Modelling & KPIs Identification

The Policy Model Editor (PME) is the component that supports and guides the policy maker (PM) to effectively model policies by selecting a data schema, applying relevant Key Performance Indicators (KPIs) or setting new ones with simple linear functions, and creating a set of rules (criteria). As for the existing policies, the PM shall name a description with a set of rules (criteria) which applies the values of a specific data schema and KPIs. The Requirements for the Policy Management Framework are provided in Deliverable D2.5 [23] (section 5.4).

Our model establishes an N-N relationship between Policies and KPIs, i.e. a single KPI can be used by many Policies and many KPIs can be included (used) by a single Policy. These relationships are realized within the PME where the end-user can associate to a single Policy multiple KPIs. KPIs can be associated with multiple domains (e.g. Security_Domain, Agrifood_Domain, Labour_Domain etc.). During the creation of a single Policy, the PME provides to the end-user the option to select multiple KPIs that are associated with the specific domain the Policy belongs to. An interesting aspect of our model and PME is the following: As KPIs can belong to multiple domains, a selected KPI (e.g. average_income_KPI) for a specific Policy (e.g. Agrifood_Policy) in a specific domain (e.g. Agrifood_Domain) may have originally been created in the Labour_Domain through a specific policy (e.g. Labour_Policy). Thus, the specific KPI in our example while originally created for the Labour_Policy/Labour_Domain, it can find application and be used in the Agrifood domain (through the Agrifood_Policy/Agrifood_Domain).

In summary, our model through the PME enables the reuse of a KPI, originally created by a specific Policy belonging to a specific domain, by another policy in another domain, enabling a "cross-over" between policies.

7.7.2 Middleware for Policies

A middleware based on .NET Core has been designed and implemented as the adapter pattern to retrieve data from the policy datastore. At the other end of the adapter lies a REST API as a mechanism that allows policies to be modelled and designed based on specific structural representations.

For more details, please refer to Deliverable D5.4 "Cross-sector Policy Lifecycle Management: Design and Open Specification 2".

7.8 Layer 4 - Policy Development Toolkit

Components: Policy Development Toolkit (T5.3), Data Visualization (T5.3)

7.8.1 Policy Development Toolkit and Data Visualization

The Policy Development Toolkit (PDT), along with the Policy Model Editor (PME), constitute the Front-End of the PolicyCLOUD platform. They integrate several sub-components to enable policy makers (PMs) to create, update and validate their policies. The PM will trigger the underlying analytics mechanisms to provide the corresponding quantitative information, while integrating the visualization component to ensure that the results are presented in a meaningful way. It includes mechanisms to explore and incorporate available analytics into new or existing policy models. The PM will set Key Performance Indicators (KPIs) that support the policy in focus, which will be calculated through the triggering of selected suitable analytics along with the provision of the respective parameters regarding datasets, temporal or spatial constraints, population filtering etc. The Requirements for the Policy Development Toolkit are provided in Deliverable D2.5 [23] (section 5.5 and section 6.14).

For the visualization of analytical tools results, the PolicyCLOUD platform provides a reporting tool that enables to build visual analytical reports. The reporting is produced from analytical queries and includes summary tables as well as graphical charts resulted from the analytics. The dashboard is adaptable, since it enables the inclusion of different charts with the KPIs chosen by the PM and a set of transformation operators that can aggregate and correlate the received policies KPIs.

The PDT directly interacts with the Data Acquisition and Analytics (DAA) Layer, the datastore and the integrated visualization as presented in the next section.

7.8.2 PDT Architecture

The present section describes the functional architecture of the Policy Development Toolkit (PDT). As a single page web application, PDT hides the complexity of the system dataflow to provide to policy makers (PMs) an integrated Decision Support System (DSS) towards the application of evidence-based Public Policies (PPs).

The general interconnection of the PDT with the other PolicyCLOUD components is illustrated in Figure 15. PDT may be considered as the point of integration and interaction of the platform with the PMs. Through the PDT, the PM is able to question the platform data and exploit the analytics tools to perform policy modelling and evaluation.

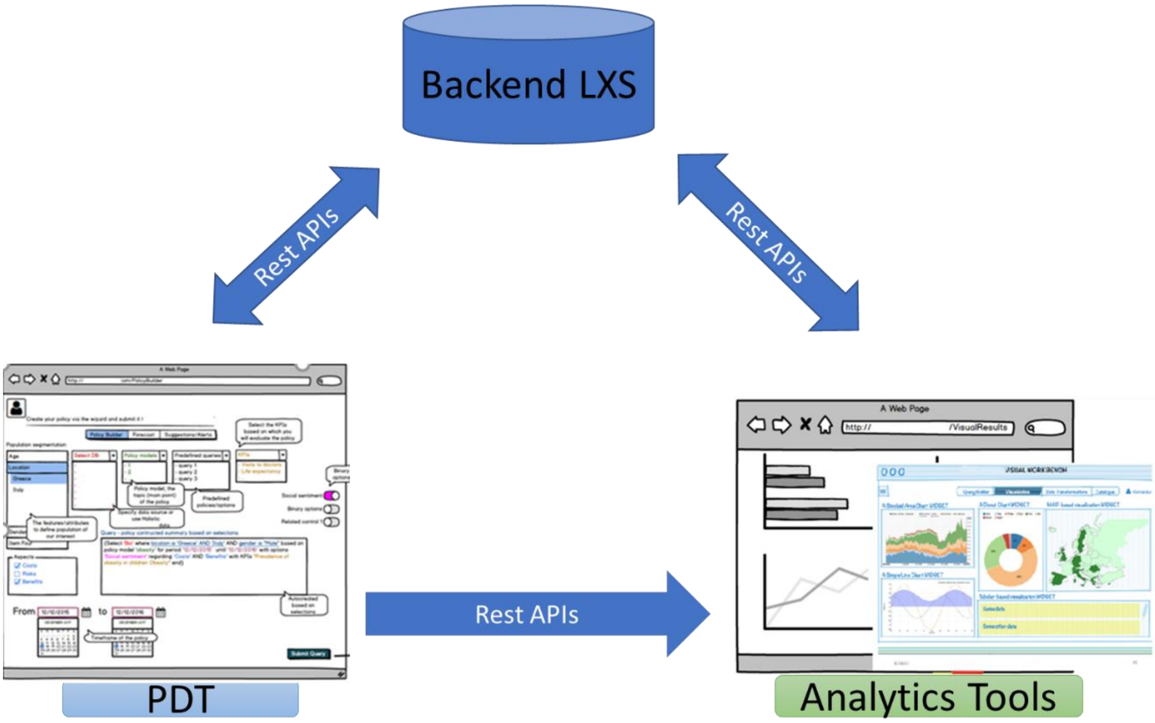


FIGURE 15 - POLICY DEVELOPMENT TOOLKIT COMMUNICATION COMPONENTS

Figure 15 shows the two main components with which PDT will communicate: Backend / Data Repository and the various Analytics Tools.

Both components will expose API Interfaces so that PDT - as the front-end UI - receives the policy model related data from datastore along with the list of registered policy-related data sources and analytic functions. It then activates the selected analytic function on a predefined data source with the parameters specified by the PM. The arrows in Figure 15 depict the communication between the components through REST APIs. The Analytics Tools become available to the PDT once they are registered to the platform. The Analytics Tools registration sequence is provided in Section 7.6

The Policies are serialized in a predefined format following common syntax (in JSON) into the datastore. The PDT translates/deserializes the policy objects retrieved from the datastore into UI objects to provide the visual environment for the policymaker actions.

The arrow between PDT and datastore also encompasses the process of semantic or rule-based reasoning and querying. Based on the process set out in T5.2, the semantic processing of emerging policies for lifecycle policy modeling is intervened, which enables the validation of the policy structure in terms of their proper construction. They also guide policymakers to choose KPIs, avoid dysfunctional policies, and provide cross-sectional policy optimization information.

In the architecture proposed in Figure 13 each component is decoupled from the others. The modular structure allows versatility and extensibility, regarding analytics tools providers, analytics frameworks, cloud providers and deployment patterns. The -also- modular UI intentionally hides the big complexity for the users, as each component is decoupled and focused on their properties and functions. So, a Policy Model is composed and supported by related KPIs, which in turn are composed of related Analytics Tools that provide their visualization graphs. The Service-Oriented Architecture (SOA) pattern is followed by requiring the components to adhere to a common communication protocol, and by exposing consistent RESTful APIs.

7.8.3 PDT Architecture Integration

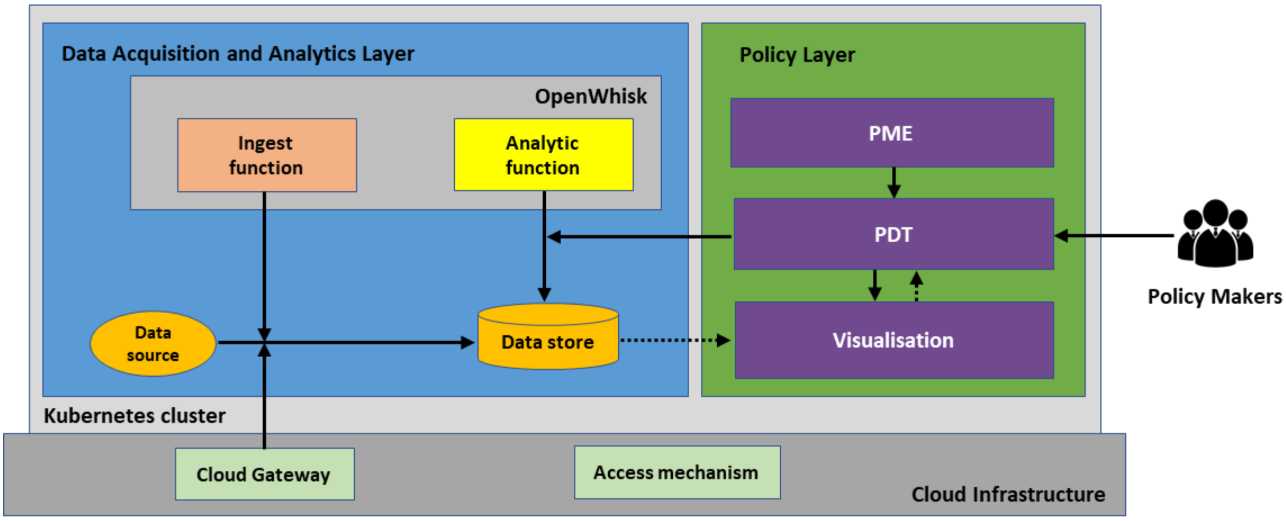


FIGURE 16 – POLICY MODEL DEVELOPMENT INTEGRATION SCHEMA

PME and PDT act in unison as the User Interface for the creation and evaluation of Policy Models (Figure 16). PME guides the user into the creation of proper policy models, while through PDT the user can call for the evaluation of policy KPIs. The analytics results are shown into the same UI depicting KPI values calculation / trends by the integrated visualization component. PME, PDT and the Visualization component share the same source code base, run as a Single Page Application (SPA), hosted on the same Web Server under the same Virtual Machine in the Cloud. All the three integrated components communicate through calls to the Rest APIs offered from the other PolicyCLOUD platform subsystems: PDT Backend, DataStore (WP4), Analytics Tools (WP4) and KeyCloak User Authentication (WP3).

Figure 17 shows the PME and PDT interaction with the PDT-Backend which provides the Restful API. Through asynchronous calls, the JSON descriptions of the available registered Analytic-Tools are retrieved from the database. These descriptions deliver the necessary information regarding the type and format of the parameters each Analytic Tools expects. The UI initializes and displays the proper interface components as to present the parameter values with the proper format (e.g. range, list, default values etc.).

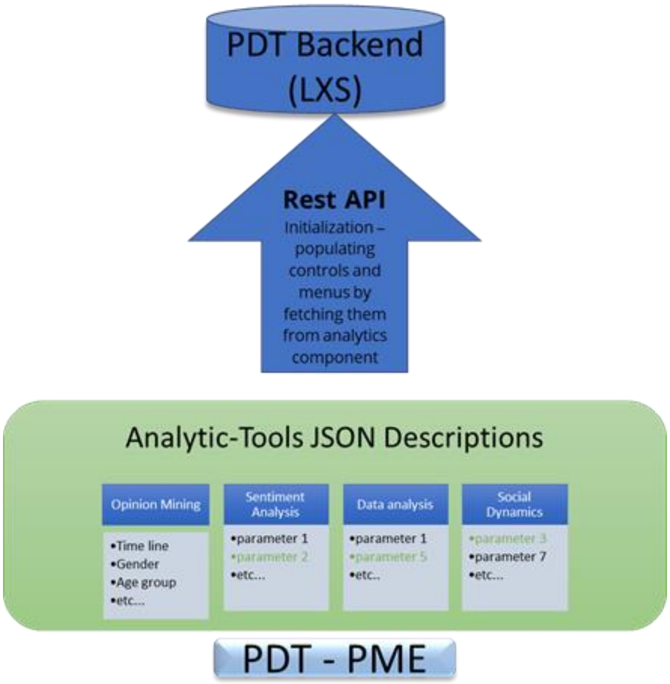


FIGURE 17 - USER INTERFACE INITIALISATION OF POLICY MODEL COMPONENTS (ANALYTIC-TOOLS)

Figure 18 depicts the retrieval process of Analytics results. The PM selects from the PDT the results that are of interest. The corresponding json objects are retrieved via the Rest API from the PDT-Backend. The Visualization Component takes over for the creation of proper graphs, according to the analytics-json object requirements.

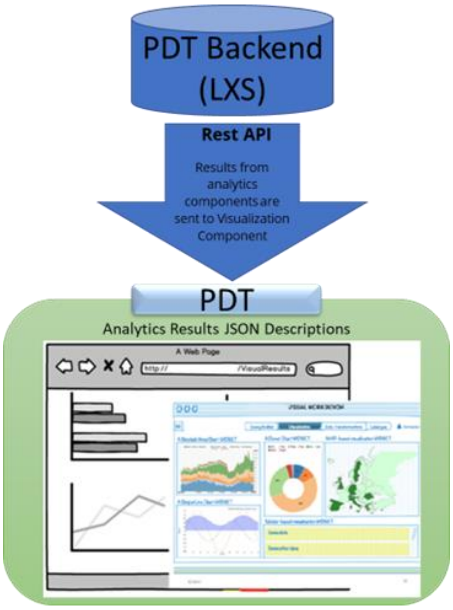

PDT Backend (LXS)
Rest API
Results from analytics components are sent to Visualization Component
PDT
Analytics Results JSON Descriptions
A Web Page
http://
/VisualResults

FIGURE 18 - RESULTS FROM ANALYTICS FETCHED FROM BACKEND AND VISUALIZED IN PDT

7.9 Layer 5

7.9.1 Data Marketplace

In the context of PolicyCLOUD, the Data Marketplace is a public, unified and standalone platform with many different APIs, able to store several types of assets (solutions). The offered assets vary depending on the needs of the project's stakeholders. They may derive/result from the separate procedures and mechanisms that are implemented in the PolicyCLOUD platform or in general, may be outcomes of the project (policies, templates, tutorials, and others), or may be even material coming from external users that is related with the overall concept and implementation of the project. The System Requirements for the Data Marketplace are provided in Deliverable D2.5 [23] (section 5.6).

From its architecture perspective, the Data Marketplace is structured around two core services, the back-end and the front-end. Generally, the marketplace supports access to its offerings to both end-users and other services (through the respective interfaces). In this context, the end-users are able to interact with the market platform through the front-end that reflects a user-friendly platform (providing the UI), while other additional services (e.g. project's services, 3rd parties) may interact directly with the back-end. This separation contributes towards the platform's enhancements in terms of functionality as well as provides additional information and capabilities.

The back-end side of the marketplace is a RESTful API and receives HTTP requests that trigger the platform's implemented functionalities. As depicted in the following figure (Figure 19), the back-end includes three layers (i.e. Assets Storage Layer, Assets Management Layer, Interaction Layer), while the front-end includes the fourth layer of the Data Marketplace (i.e. the Presentation Layer) that in full consists of all of these 4 different layers. Their capabilities are shortly described below: The "Assets Storage Layer" is the layer in which the platform's offered assets are stored.

- The "Assets Management Layer" delivers all the needed principles and techniques for the management of the marketplace's assets.
- The "Interaction Layer" supports the communication between the marketplace and the end-users (i.e. human end-users, machine end-users), by providing discrete APIs for exploiting each different type of asset.
- The "Presentation Layer" (i.e. the front-end) provides the User Interface (UI) towards the different types of end-users that are willing to use the platform. As a result, the end-users of the Data Marketplace are able to interact with it through this layer, from which HTTP requests are sent to the "Interaction Layer". These interactions may include various requests, such as searching, creating, retrieving, updating, and deleting assets.

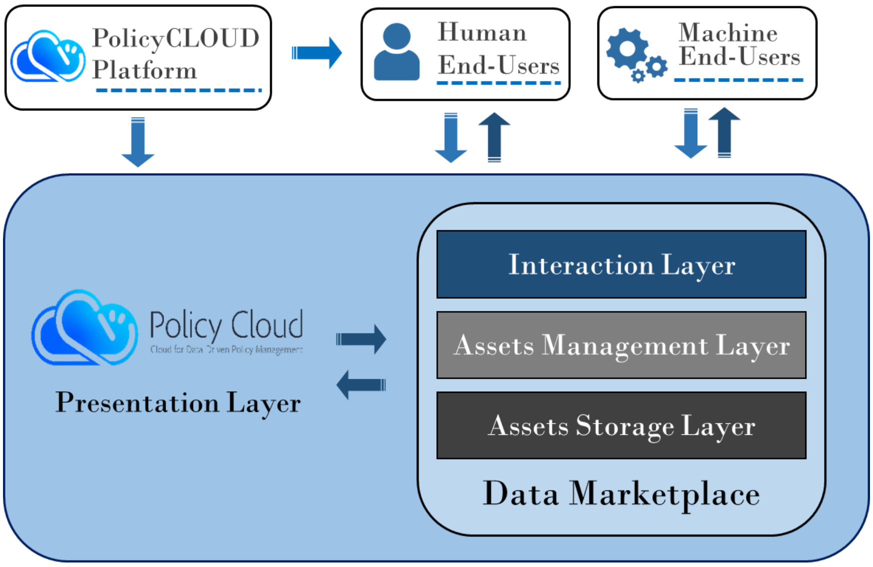


FIGURE 19 – DATA MARKETPLACE ARCHITECTURE

Figure 20 demonstrates in deeper detail the overall conceptual view which includes the three discrete layers (i.e., Assets Storage Layer, Assets Management Layer, Interaction Layer), and the front-end as the fourth layer of the Data Marketplace (Presentation Layer).

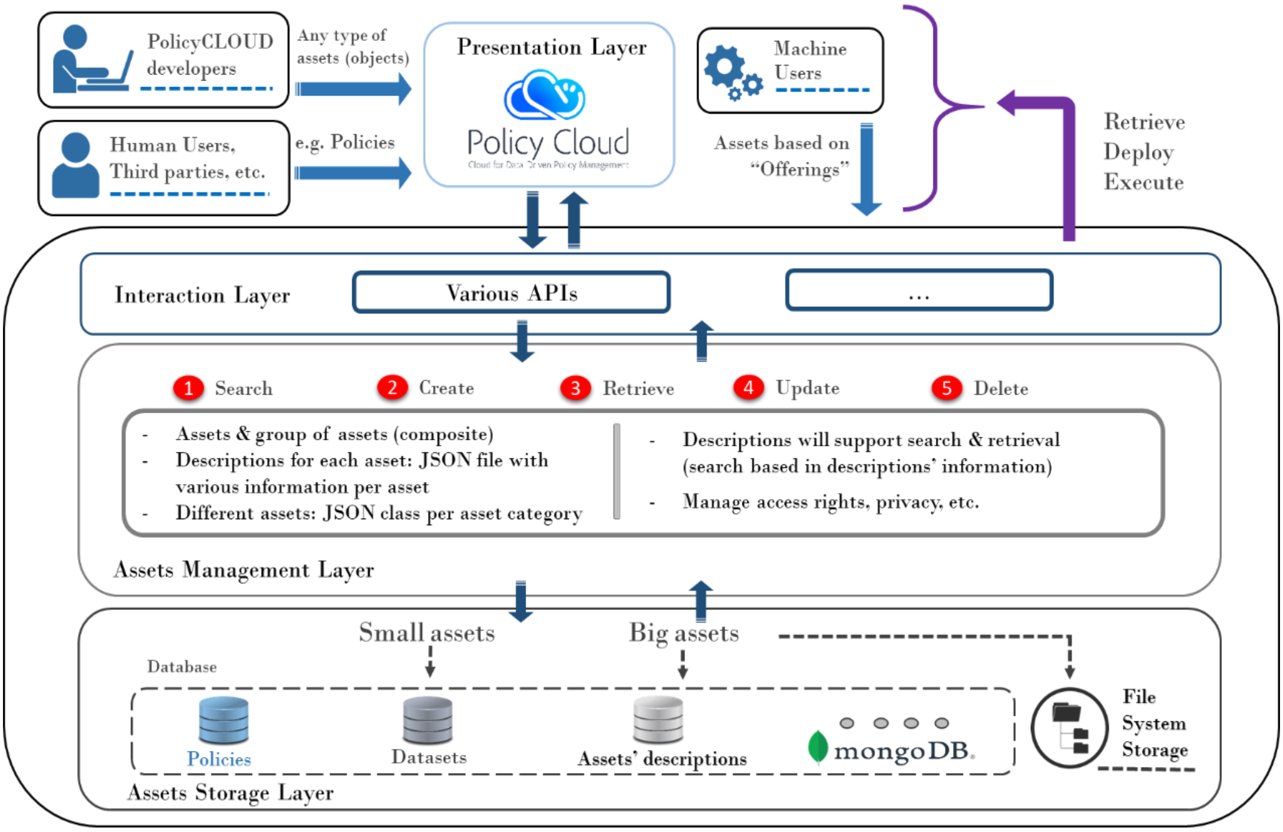

PolicyCLOUD developers
Any type of assets (objects)
Human Users, Third parties, etc.
e.g. Policies
Presentation Layer
Policy Cloud
Machine Users
Assets based on "Offerings"
Retrieve
Deploy
Execute
Interaction Layer
Various APIs
...
1 Search
2 Create
3 Retrieve
4 Update
5 Delete
- Assets & group of assets (composite)
- Descriptions for each asset: JSON file with various information per asset
- Different assets: JSON class per asset category
- Descriptions will support search & retrieval (search based in descriptions' information)
- Manage access rights, privacy, etc.
Assets Management Layer
Small assets
Big assets
Database
Policies
Datasets
Assets' descriptions
mongoDB.
File System Storage
Assets Storage Layer

FIGURE 20 - DATA MARKETPLACE: OVERALL CONCEPTUAL VIEW

7.10 Data Governance Model, Protection and Privacy Enforcement

Components: Access Mechanisms (T3.6)

7.10.1 Data Governance Model, Protection and Privacy Enforcement

The data governance model and the tools for protection and privacy enforcement are used to protect data and ensure decisions across the complete path following specific guidelines and legislations. Data Governance Model and Privacy Enforcement mechanism is depicted vertically in the right part of the Overall Architecture in Figure 4. This includes three different parts, a) the access policy editor, b) the model and model editor and c) the ABAC authorization engine. The access policy editor will provide the user with the ability to define and store policies based on the ABAC scheme according to the XACML standard. The data governance model of PolicyCLOUD will be used for the definition of these policies, and also for the actual enforcement of the policies by the authorization engine that will be able to evaluate the policies and the attributes, thus enforcing protection and privacy-preserving policies.

In addition, as depicted in Figure 4 and presented for convenience in Figures 21 (A), (B) and (C), the components developed in the scope of T3.6 regarding the protection of data and privacy enforcement, will be used in three separate parts of the overall architecture envisioned for the PolicyCLOUD. The first - Figure 21 (A) - is to provide an access control mechanism for the inclusion and usage of data sources that are being part of PolicyCLOUD. The second - Figure 21 (B) - is the access control being also provided at the level of data visualization, thus allowing or denying access to specific data analytics. The third - Figure 21 (C) - is the usage of the access control mechanisms for managing the control between the PolicyCLOUD datastore and any additional private data store that may be used.

Finally, for the whole mechanism to work properly it has to be mentioned that the authorization engine will need to have access to the attribute values regarding the data, the data sources/origins, the phase of the data lifecycle (e.g. stored data or analysed data) and the phase of the policy lifecycle (e.g. modelling or experimentation process); these can be provided by external components acting as adapters, and can be developed per use case.

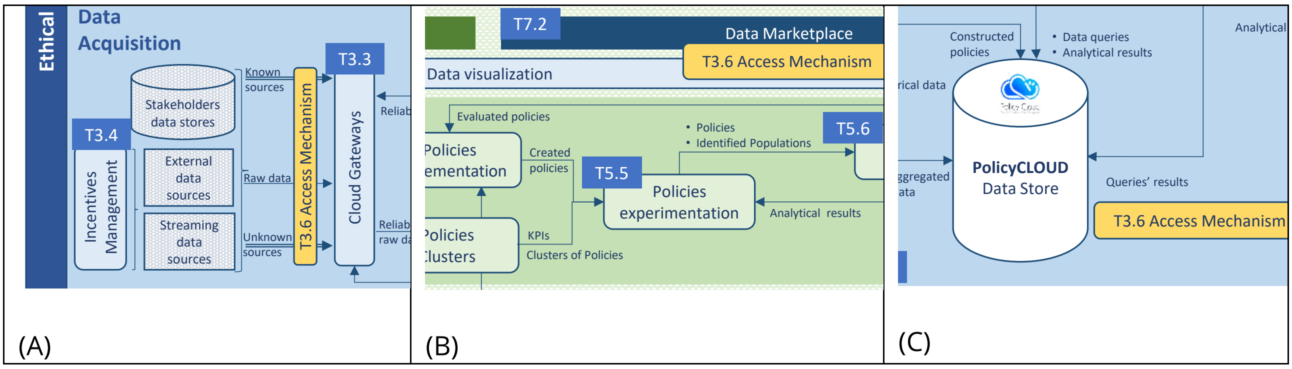


FIGURE 21 – DATA GOVERNANCE MODEL, PROTECTION AND PRIVACY ENFORCEMENT MECHANISMS – EXTRACTED VIEWS (A), (B) AND (C) FROM THE DIAGRAM OF POLICYCLOUD OVERALL ARCHITECTURE.

7.10.2 Data Governance model, protection and privacy enforcement mechanisms Integration

7.10.2.1 Integration with the Overall Architecture

The integration of the Data Governance and Privacy Enforcement Mechanism is achieved via two components, the Keycloak and the ABAC servers that are connected to each other. As presented in Figure 22, the pair of Keycloak and ABAC can intercept requests to both the Policy Development Toolkit and the Cloud Gateways, ensuring the privacy enforcement for both. A client makes a request to either of those services (STEP 1) and is immediately intercepted by the ABAC Engine. In order to make a decision, ABAC queries the Keycloak server regarding the attributes of the user making the request (STEP 2). Keycloak provides the requested user attributes (STEP 3) and depending on whether they satisfy the current implemented ABAC policy the request is allowed or denied to go through to its original destination (STEP 4). Currently the mechanism has been integrated with the Policy Development Toolkit (PDT), the Cloud Gateways and the Marketplace. Furthermore, Keycloak has been integrated with EGI Check-In, in order for the platform to provide additional login options via academic or social credentials. This interaction is presented on the top left corner of Figure 22.

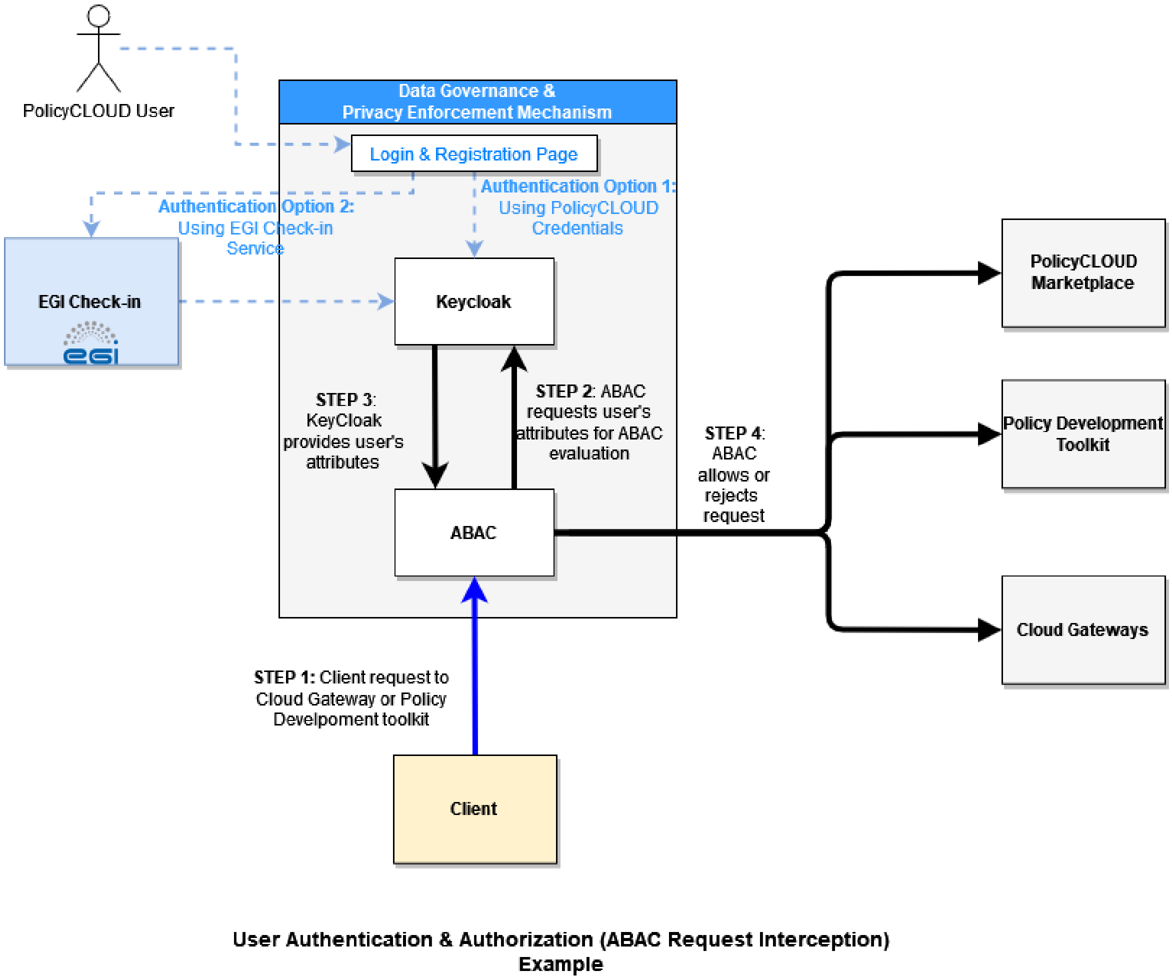

PolicyCLOUD User
Data Governance & Privacy Enforcement Mechanism
Login & Registration Page
Authentication Option 1: Using PolicyCLOUD Credentials
Authentication Option 2: Using EGI Check-in Service
EGI Check-in
egi
Keycloak
STEP 3: KeyCloak provides user's attributes
STEP 2: ABAC requests user's attributes for ABAC evaluation
ABAC
STEP 4: ABAC allows or rejects request
PolicyCLOUD Marketplace
Policy Development Toolkit
Cloud Gateways
STEP 1: Client request to Cloud Gateway or Policy Develpoment toolkit
Client
User Authentication & Authorization (ABAC Request Interception) Example

FIGURE 22 - DATA GOVERNANCE & PRIVACY ENFORCEMENT MECHANISM INTEGRATION FLOW

7.10.2.2 INTEGRATION WITH THE KUBERNETES CLUSTER

The components of the architecture described in the previous sections have been deployed through EGI provisioned infrastructure and have thus been integrated to the Kubernetes cluster. More specifically, an instance of the Keycloak server, along with a connected instance of the ABAC Server have been made available.

8 Use Case examples for end-to-end data path analysis

In the following sections a short description of the Use Case scenarios which includes problem statement, main objectives, Key Performance Indicators and data sources to be used, is provided (sections 8.1, 82, 8.3 and 8.4) based on deliverables D6.10 [25] , D6.11 [24] and D2.5 [23].

The first scenario (scenario A for Use Case 1 "Participatory policies against Radicalization"), serves as an end-to-end example while additional scenarios from the different Use Cases follow a similar end-to-end data path, supported by the generality of the PolicyCLOUD environment (section 8.7), demonstrating the data ingest flow and data exploitation while analysing the processing and data transformations along the complete data path.

From the technical perspective an end-to-end data path analysis is provided through the integration of two subpaths: (i) the path from the Data Sources-Cloud Gateway to the LXS database, which is analyzed in section 8.5 and (ii) the path from the LXS database-PDT backend to the Visualization-PDT interface with which the Policy Maker (as end-user) interacts, analyzed in section 8.6. The two subpaths constitute a complete end-to-end data path from a data source to a semantically meaningful result to be presented to the end user.

8.1 Use Case 1: Participatory Policies Against Radicalization

8.1.1 Scenario A: Radicalization incidents

Description:

Monitor the occurrence of radicalization incidents in the geographic proximity of a region. Data coming from the GTD and RAND will be used. The Policy Maker can select the area of his/her interest and consult the different incidents that have taken place in a given period.

Detailed description of the scenario and user/stakeholder requirements are included in Deliverable D2.5 [23] (sections 2.1.1-2.1.3) and in Deliverable D6.11 [24] (sections 2.1.1-2.1.3).

8.1.2 Main Objective

Validate existing policies and investigate if there is a need to update them or create new ones based on the retrieved information.

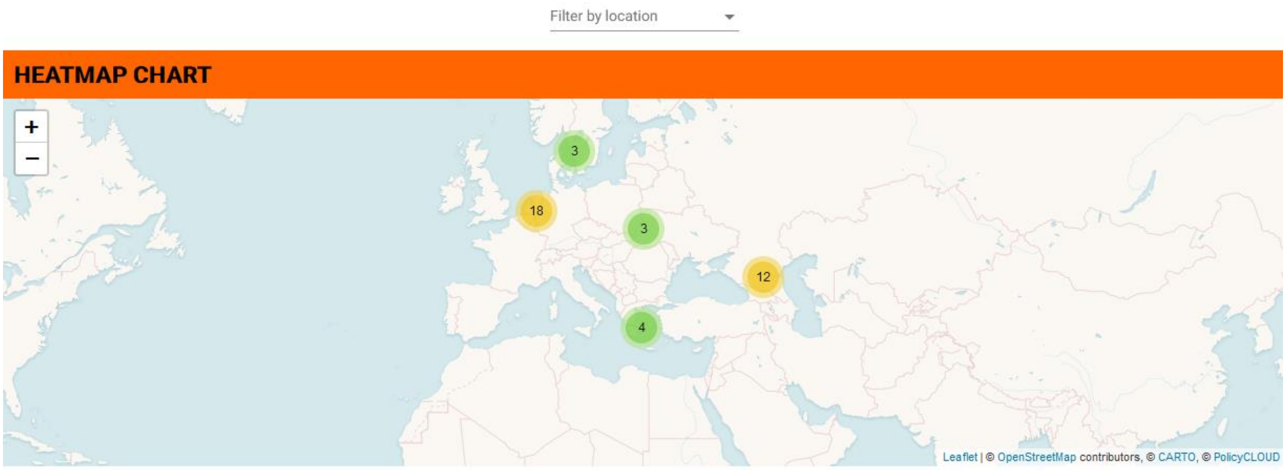

Filter by location
HEATMAP CHART
+
−
3
18
3
12
4
Leaflet | © OpenStreetMap contributors, © CARTO, © PolicyCLOUD

FIGURE 23 - VISUALIZATION ON POLICYCLOUD OF THE RESULT OF SCENARIO A: RADICALIZATION INCIDENTS OF USE CASE 1

8.1.3 Key Performance Indicators

Section	Description
ID	MAG-KPI7 (D6.11, Table 14)
Title	Number of identified occurrences of radicalization incidents in a given area
Priority	High
Reference Use Case	UC#1
Success Criteria	>=0

TABLE 1 – UC1 BUSINESS KPI7

8.1.4 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Participatory Policies Against Radicalization	Scenario A	Managed by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), the Global Terrorism Database includes more than 200,000 terrorist attacks dating back to 1970.	https://www.start.umd.edu/gtd/access/

TABLE 2 – PARTICIPATORY POLICIES AGAINST RADICALIZATION USE CASE, SCENARIO A - DATA SOURCES LIST

8.2 Use Case 2: Intelligent policies for the development of the agrifood industry

8.2.1 Scenario B: Visualization of negative and positive opinions on social networks for different products

<u>**Description:**</u>

Visualize the negative and positive opinions on social networks of the different products analysed allowing an automatic and immediate response to the end user.

Detailed description of the scenario and user/stakeholder requirements are included in Deliverable D2.5 [23] (sections 2.2.1-2.2.3) and Deliverable D6.11 [24] (sections 3.1.1-3.1.3).

8.2.2 Main Objective

Create an immediate communication with the end user, knowing their impressions, both positive and negative, that will allow us to interact with the end customer more directly.

8.2.3 Key Performance Indicators

Section	Description
ID	SAR-KPI4 (D6.11, Table 26)
Title	Provide real–time calculation capacity
Priority	High
Reference Use Case	UC#2
Success Criteria	>20% of the data

TABLE 3 – UC2 TECHNICAL KPI4

Section	Description
ID	SAR-KPI6 (D6.11, Table 28)
Title	Increase process speed
Priority	High
Reference Use Case	UC#2
Success Criteria	>30% Reduce time

TABLE 4 – UC2 TECHNICAL KPI6

Section	Description
ID	SAR-KPI8 (D6.11, Table 30)
Title	Total number of occurrences
Priority	High
Reference Use Case	UC#2
Success Criteria	>50%

TABLE 5 – UC2 BUSINESS KPI8

Section	Description
ID	SAR-KPI9 (D6.11, Table 31)
Title	Relative Total nº occurrences %
Priority	High
Reference Use Case	UC#2
Success Criteria	>10%

TABLE 6 – UC2 BUSINESS KPI9

Section	Description
ID	SAR-KPI10 (D6.11, Table 32)
Title	Opinion (-1 (negative) to 1 (positive)) |impact
Priority	High
Reference Use Case	UC#2
Success Criteria	Average positive

TABLE 7 – UC2 BUSINESS KPI10

Section	Description
ID	SAR-KPI11 (D6.11, Table 33)
Title	Increment of the impact in the last month
Priority	High
Reference Use Case	UC#2
Success Criteria	>15%

TABLE 8 – UC2 BUSINESS KPI11

8.2.4 Data Sources

Link #	Link
1	https://opendata.aragon.es/datos/catalogo?texto=pac
2	https://www.aragon.es/en/-/vitivinicultura.-registro-viticola
3	https://www.aragon.es/en/temas/medio-rural-agricultura-ganaderia/agricultura/vinedos-vinos-bebidas-alcoholicas
4	https://opendata.aragon.es/datos/catalogo/busqueda/siu?tema=vinedos-vinos-bebidas-alcoholicas
5	https://opendata.aragon.es/servicios/open-social-data/#/main

TABLE 9 – LINKS TO ARAGON USE CASE DATA STORES

8.3 Use Case 3: Facilitating urban policy making and monitoring through crowdsourcing data analysis

8.3.1 Scenario A: Road Infrastructure - Visualization of signals received from Sofia's Call Centre 'CallSofia'

<u>**Description:**</u>

Road infrastructure is one of the most important and budget consuming elements within the context of the urban environment that impacts citizens' everyday life. Reliable analysis is needed on current situation in all 24 district administrations, in order to foresee and improve long term policy making in the area of road infrastructure.

Detailed description of the scenario and user/stakeholder requirements are included in Deliverable D2.5 [23] (section 2.3.2 – Table 51) and Deliverable D6.11 [24] (section 4.1.3 – Table 38).

8.3.2 Main Objective

The main objectives of this scenario are: to improve long term policy making in the area of road infrastructure, and to envision and build the capacities of district administrations and municipal administration in solving road infrastructure problems.

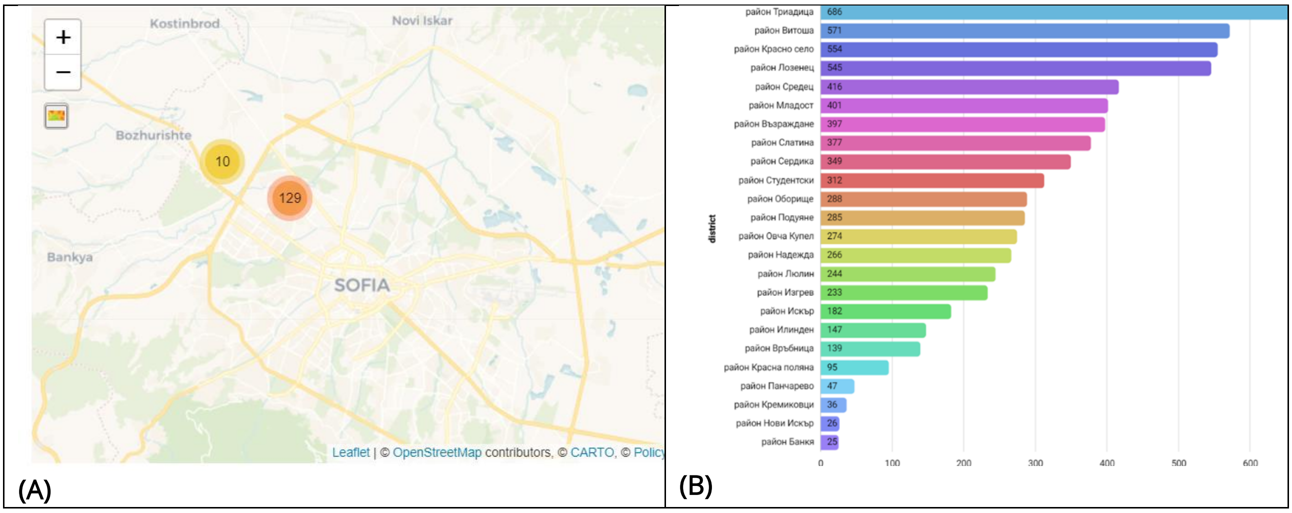


FIGURE 24 - VISUALIZED RESULTS ON POLICYCLOUD, IN THE FORM OF HEATMAP (A) AND BAR CHART (B) FOR ROAD INFRASTRUCTURE (SOURCE D6.11 [24]).

8.3.3 Key Performance Indicators

Section	Description
ID	SOF-KPI1 (D6.11, Table 43)
Title	Increased efficiency: Reduction of time to develop a policy
Priority	High
Reference Use Case	UC#3
Success Criteria	>50% of the data

TABLE 10 – UC3 BUSINESS KPI1

Section	Description
ID	SOF-KPI2 (D6.11, Table 44)
Title	Increased efficiency: Reduction of time to develop a policy
Priority	High
Reference Use Case	UC#3
Success Criteria	>20%

TABLE 11 – UC3 BUSINESS KPI2

Section	Description
ID	SOF-KPI4 (D6.11, Table 46)
Title	Increase in local ecosystem and community engagement and collaboration in urban policy development
Priority	High
Reference Use Case	UC#3
Success Criteria	>15%

TABLE 12 – UC3 BUSINESS KPI4

Section	Description
ID	SOF-KPI5 (D6.11, Table 47)
Title	Number of data sources integrated and linked to the PDT
Priority	High
Reference Use Case	UC#3
Success Criteria	>=2

TABLE 13 – UC3 TECHNICAL KPI5

Section	Description
ID	SOF-KPI6 (D6.11, Table 48)
Title	Increased speed of access to information
Priority	High
Reference Use Case	UC#3
Success Criteria	>15%

TABLE 14 – UC3 TECHNICAL KPI6

8.3.4 Data Sources

#	Data Source
1	Call Centre of Sofia Municipality 'CallSofia' (https://call.sofia.bg/)

TABLE 15 – SOFIA USE CASE, SCENARIO A, ROAD INFRASTRUCTURE - DATA SOURCE

8.4 Use Case 4: Predictive analysis towards unemployment risks identification and policy making

8.4.1 Scenario A: Unemployment Analysis

Description:

Unemployment is expected to go up during the years following the pandemic. Analysis will be performed for specific time periods in the past. Through such analysis, the expected unemployment rate following a second wave of infections can be estimated for the subsequent years.

Detailed description of the scenario and user/stakeholder requirements are included in Deliverable D2.5 [23] (section 2.4.2 – Table 56) and Deliverable D6.10 [25] (section 4.1.3 – Table 40).

8.4.2 Main Objective

The objective of this scenario is to use the analytics and visualizations produced from the PolicyCLOUD platform to identify key information that could help determine groups of citizens that are affected by unemployment.

8.4.3 Key Performance Indicators

Section	Description
ID	LON-KPI1 (D6.10, Table 42)
Title	Count of unemployed citizens under 25. This KPI will include a total count of all the citizens that are unemployed which are aged below 25.
Priority	N/A
Reference Use Case	UC#4
Success Criteria	Analytics based on the requested KPI should be clearly displayed in a number format or visualisation.

TABLE 16 – UC4 POLICY KPI1

Section	Description
ID	LON-KPI2 (D6.10, Table 43)
Title	Count of unemployed divided by age group. This KPI will include a total count of all the citizens and categorise them into various age ranges i.e. 25-40.
Priority	N/A
Reference Use Case	UC#4
Success Criteria	Analytics based on the requested KPI should be clearly displayed in a number format or visualisation.

TABLE 17 – UC4 POLICY KPI2

Section	Description
ID	LON-KPI4 (D6.10, Table 45)
Title	Annual percentage increase/decrease of females claiming benefits. This KPI will include a total count of all the citizens and categorise them by gender.
Priority	N/A
Reference Use Case	UC#4
Success Criteria	Analysis results

TABLE 18 – UC4 POLICY KPI4

Section	Description
ID	LON-KPI7 (D6.10, Table 48)
Title	User Engagement. Numerical statistics that will track the number of users that engage with the platform.
Priority	N/A
Reference Use Case	UC#4
Success Criteria	Statistics based on the amount of user's engagement with the platform should be clearly displayed in a number format or visualisation.

TABLE 19 – UC4 TECHNICAL KPI7

8.4.4 Data Sources

#	Data Source
1	Camden open data (https://opendata.camden.gov.uk/)

TABLE 20 – LONDON USE CASE, SCENARIO A, UNEMPLOYMENT ANALYSIS - DATA SOURCE

8.5 Data Path Analysis (From Cloud Gateways to LXS Database) based on the implemented Use Case scenarios

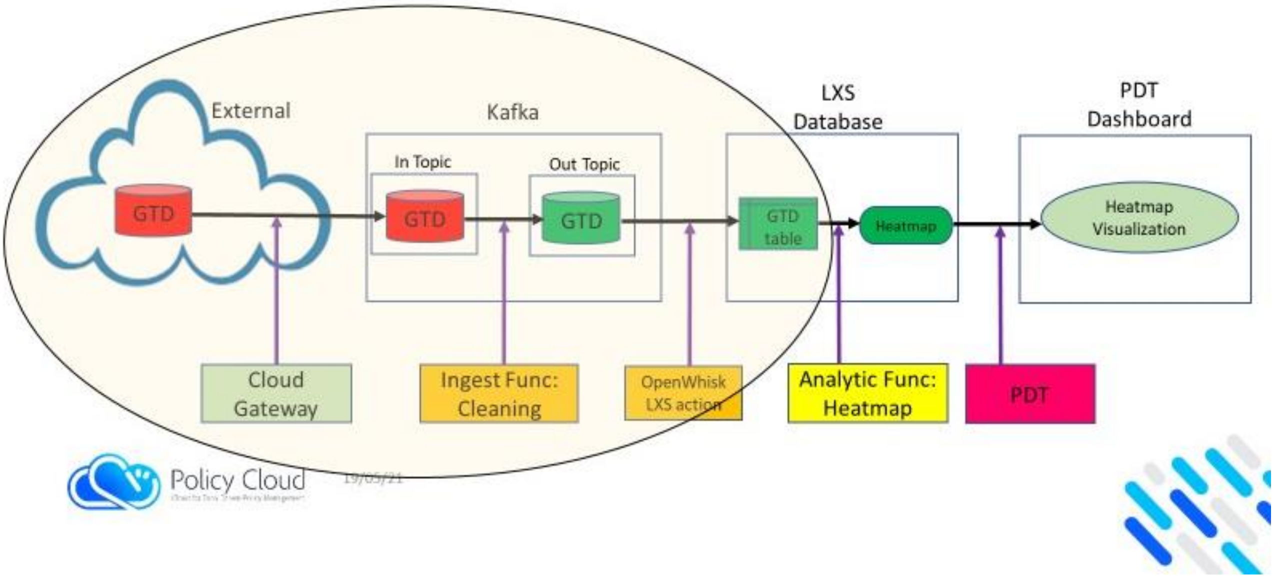


FIGURE 25 - DATA PATH ANALYSIS

Data Path Analysis will be performed for the first scenario (scenario A) of Use Case 1 “Participatory policies against Radicalization”. The scenario serves as an example, while additional scenarios from the different Use Cases follow a similar end-to-end data path, supported by the generality of the PolicyCLOUD environment (section 8.7).

More specifically, scenario A of Use Case 1, provides to a Policy Maker a visualization of a heatmap showing the frequency of occurrence of radicalization incidents in the geographic proximity of a region. Data coming from the GTD is used. Figure 25 demonstrates the data path for this end-to-end example, for the function implementing the heat map computation which is invoked when the heatmap visualisation is called.

The same architecture shown in Figure 25, is also used for the first scenario (scenario A) of Use Case 2 “Intelligent policies for the development of agrifood industry” which provides to a Policy Maker a visualization of the ARAGON wine sentiments with data received from social media. It could be mentioned that the social media (ARAGON wine sentiments) results can also be presented in the same manner as the results from the GTD.

8.6 Data Path Analysis (from the LXS database backend to visualization of result)

Data Path Analysis highlights the integration among the various components that consist the PDT on the one hand, and the corresponding building blocks of the overall PolicyCLOUD architecture on the other hand, Figure 2626 provides the sequence diagram of all interactions that take place when a PM invokes an analytical function and receives the results in a visualized graph.

As depicted in the sequence diagram, the end user of the PDT, the Policy Maker, retrieves all existing policies stored in the platform, according to some filter criteria. The PDT returns these policies and visualizes them in its graphical user interface. Then, the Policy Maker wants to verify a specific policy, by making an analysis over the available data. Using the graphical user interface of the PDT, it selects and clicks on the KPI that the Policy Maker wants to verify. The GUI invokes the corresponding REST web method of the backend to execute the relevant analytical function.

Subsequently, the backend **interacts with the data acquisition and analytics layer** of the PolicyCLOUD (sections 7.6.11.1 and 8.3). It contains all the information regarding registered analytical functions, their required input parameters, the type of their output etc. As a result, it collects the data received by the invocation of its REST web method, and further requests from the DAA layer to execute the function.

As explained in the previous subsection, the DAA layer incorporates the OpenWhisk serverless platform, and relies on the latter to deploy the requested function. Openwhisk takes the responsibility to create, via Kubernetes, the corresponding infrastructure resources, deploy its runtime execution environment there and finally executes the requested function.

The function on the other hand, receives its required input parameters that were passed to it via the PDT, the backend and the DAA layer, along with other meta-information (such as the connection URL of the data management layer to retrieve data). Through this process it receives all required information to open a database connection with the datastore, and execute its relevant query, thus pushing a pre-processing down to the storage layer, in order to retrieve only the amount of data that is needed to run its AI algorithm.

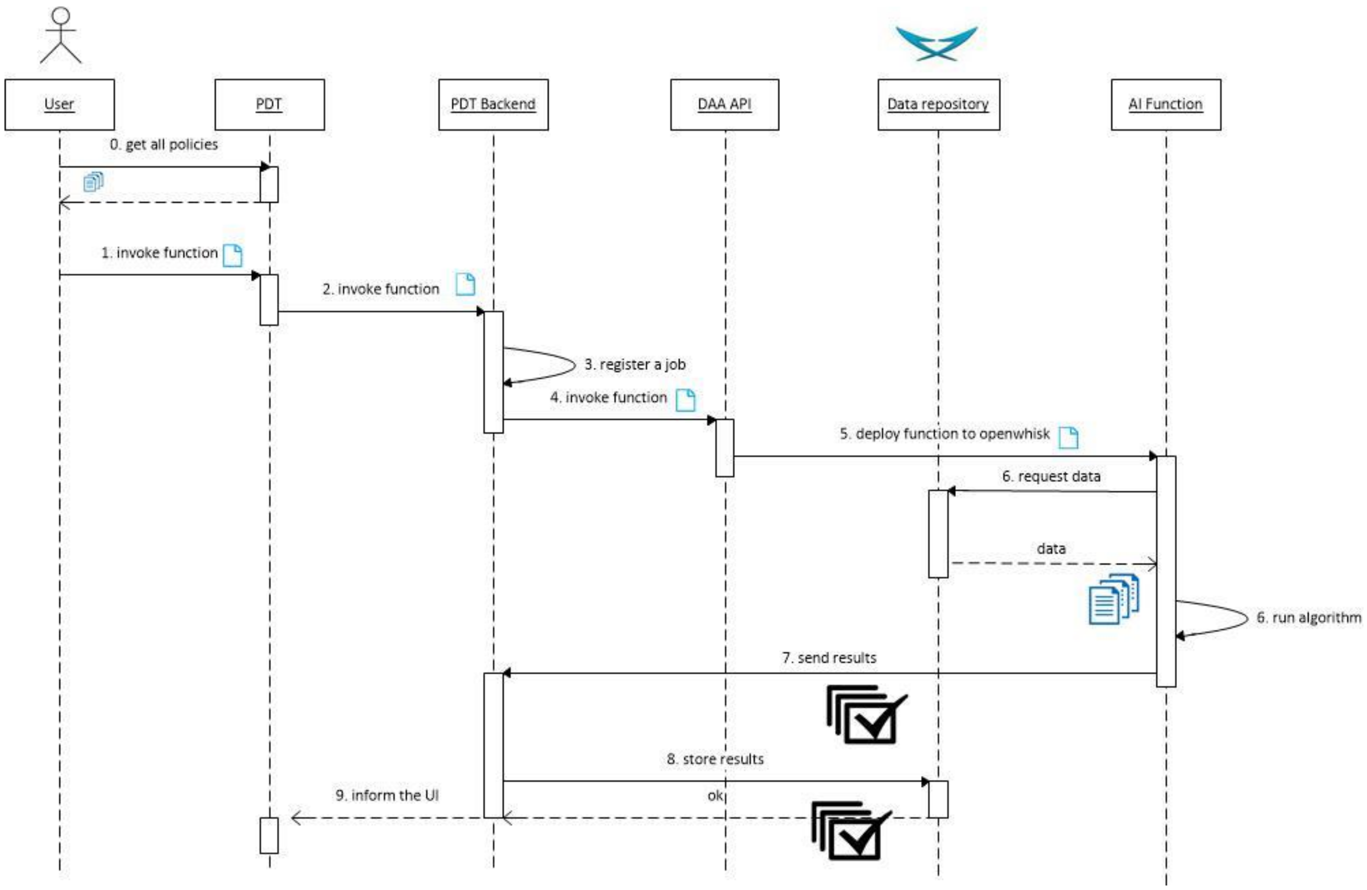


FIGURE 26 - SEQUENCE DIAGRAM FOR PDT-DAA INTERACTION

Among other meta-information parameters received, one important parameter is the URL of a REST web service that the function needs to communicate to persistently store the results of the analysis. In fact, when the AI algorithm produces results, and before the function completes, it firstly sends the results in this URL provided by the PDT backend, so that the latter can store this information and make it available to the end-user, the Policy Maker. After sending the results, the function returns, so that the Openwhisk can shut down the relevant run-time environment and release the resources used. When the function has been properly shut down, it informs the DAA layer, and the result can be further communicated to the PDT backend. Finally, when the REST web service of the backend is invoked in order to store the results, the backend persistently stores them into the data repository, by adding relevant meta-information. At the end of this process, it informs the PDT via web sockets that the results are now available and can be retrieved. The PDT sends a pop-up notification in the graphical user interface, and the Policy Maker can now click and see the results. The PDT will get the results upon request from the backend, and it will activate and make visible the corresponding type of visualization graph to show these results.

8.7 PolicyCLOUD extensibility, generality and scalability

PolicyCLOUD architecture is extensible as it supports the plugin of additional analytical tools. More specifically, the Data Acquisition and Analytics layers of the PolicyCLOUD architecture provide extensibility and reusability of analytic functions. New analytics functions (services) can be registered into PolicyCLOUD and reused for applying analytics on new and existing registered data sources.

PolicyCLOUD architecture allows all involved components to be generic enough and not locked into a specific deployment or implementation. For instance, the analytical functions do not need to know in advance where to connect in order to retrieve data, or what is the schema of the underlying data, or where to store results. They can be generic and receive this information at runtime. In the same sense, while the DAA layer has been developed to cover the needs of the PolicyCLOUD platform, it is generic and can be used by any application that needs to access programmatically a serverless platform in order to administrate, deploy and execute functions over this environment. In the same direction, the PDT backend provides an interface for different connectors. An implementation of these connectors has been developed to allow the integration of the PDT with the DAA layer. In that sense, the PDT is not locked-in to a specific platform, but it can use any other type of environment by implementing the relevant connector.

Since the involved components are generic, individual components can increase their sustainability, by being exploitable in other deployments or other integrated solutions. The analytical provider does not need to implement its function specifically for the PolicyCLOUD environment only, rather can he or she only focus on the AI algorithm, which can be used in different solutions. In the same manner, the DAA layer does not provide capabilities specifically for the PolicyCLOUD environment only, but it can be exploited in any integrated solution that requires to programmatically administrate the serverless platform. Finally, the PDT can be deployed in other environments that might not allow for dynamic deployments. However, by putting all these layers together integrated into the PolicyCLOUD, the platform can benefit from all the advancements provided by the individual components, and this is what makes the overall integrated platform so innovative.

PolicyCLOUD architecture, as a cloud architecture deployed on a Kubernetes cluster provisioned through EGI (RECAS-BARI), is highly scalable. Seamless analytics (presented in section 7.6) permits storage scalability, while the serverless OpenWhisk also permits scalability both for ingestion and for analytics on data at rest.

In summary, during the project lifespan the PolicyCLOUD platform has been successfully exploiting cloud resources (i.e. Infrastructure as a Service). The key building blocks of the platform are using cloud-native technologies. Due to the elastic nature of the cloud, the PolicyCLOUD platform is capable of scaling up to meet the ever increasing demands of data-driven policy-making initiatives.

9 Conclusion

The third and final version of the PolicyCLOUD Conceptual Model & Reference Architecture (Deliverable D2.7, originally submitted as Deliverable D2.2 with the second version submitted as Deliverable D2.6) is presented in this document.

PolicyCLOUD architecture, is an extensible, highly scalable cloud architecture supporting generic components. The architecture consists of the following five layers: Cloud Based Environment (Layer 1a), Data Management – Data Stores (Layer 1b), Data Acquisition and Analytics (Layer 2), Policies Management Framework (Layer 3), Policy Development Toolkit (Layer 4) and Data Marketplace (Layer 5). The architecture also includes the Ethical Framework and the Data Governance Model, Protection and Privacy Enforcement.

Special emphasis has been given on the Integration in PolicyCLOUD which follows three directions: (i) architecture integration, (ii) integration with the cloud infrastructure and (iii) integration with Use Case scenarios through the implementation of end-to-end scenarios.

Use Case scenarios are used for end-to-end data path analysis: The data path analysis consists of two subpaths: (i) the subpath from the Cloud Gateways to LXS database and (ii) the subpath from the LXS database backend to the visualization of result. The two subpaths constitute a complete end-to-end data path from an external data source to a semantically meaningful result to be presented to the end user.

Additional integration activities took place along the two frameworks of PolicyCLOUD, (a) the Data Governance model, protection and privacy enforcement mechanism and (b) the Ethical and Legal Compliance framework.

This final version of the document provides additional information on how External Frameworks can be integrated with PolicyCLOUD. It also provides the overall Conceptual View and architecture of the Data Marketplace. The mechanisms developed for initialising the Policy Development Toolkit with Policy Model components and the visualization of results are also presented.

The document also addresses the Reviewers' comments for the previous version of the deliverable (Deliverable D2.6), which were included in the second review report. More specifically the updates in Deliverable D2.7 also include: (i) links to specific user/stakeholder requirements (D2.5), (ii) descriptions and implementation details for the two remaining pilot Use Cases (Sofia and London) - sections 8.3 and 8.4 and (iii) reference to EOSC (section 7.3.3) and to the role of the Conceptual Model & Reference Architecture document for the identification of the relevant services and of their providers, and description of the onboarding process based on Deliverable D3.4.

Within the context of the Ethical and Legal Compliance Framework positive interventions to the PolicyCLOUD architecture are introduced, including the addition of specific fields/parameters to the registration Application Programming Interfaces to be populated with details regarding each individual analytics tool and dataset/data source.

A new feature outlined in the document is also the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Policy Development Toolkit, the cloud gateways and the marketplace. Within the same context, the integration of EGI-Check-in with Keycloak including the integration of the Data Governance model, protection and privacy enforcement mechanisms with the Kubernetes cluster are presented.

References

[1] Apache Mesos, “http://mesos.apache.org/”

[2] IBM Cloud Object Storage, https://www.ibm.com/cloud/object-storage

[3] PolicyCLOUD. D3.1 - Cloud Infrastructure Incentives Management and Data Governance Design and Open Specification 1. Ledakis, Giannis. 2020.

[4] Load balancing, Moleculer “https://moleculer.services/docs/0.14/balancing.html”

[5] Fault tolerance, Moleculer “https://moleculer.services/docs/0.14/fault-tolerance.html”

[6] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A distributed messaging system for log processing", NetDB, 2011.

[7] Malone, Thomas W., Robert Laubacher, and Chrysanthos Dellarocas. "The collective intelligence genome." MIT Sloan Management Review 51, no. 3 (2010): 21.

[8] PolicyCLOUD. D3.4 - Cloud Infrastructure Incentives Management and Data Governance Design and Open Specification 2. Ledakis, Giannis. 2021.

[9] PolicyCLOUD. D3.3 - PolicyCLOUD’s Societal and Ethical Requirements & Guidelines, Audino Alice, 2020.

[10] Gehra B., Leiendecker J. & Lienke G. (2017), White Paper. Compliance by Design: Banking’s Unmissable Opportunity, “https://image-src.bcg.com/Images/Compliance-by-Design-Dec2017_tcm9-198779.pdf” retrieved 2020-11-28.

[11] Balboni P., Taborda Barata M., Botsi A. & Francis K. (2019), Accountability and Enforcement Aspects of the EU General Data Protection Regulation – Methodology for the Creation of an Effective Compliance Framework and a Review of Recent Case Law, Indian Journal of Law and Technology, 15(1), 102-259.

[12] JSON, “http://json-ld.org/”

[13] Yamada, I., & Shindo, H. (2019). Neural attentive bag-of-entities model for text classification. *arXiv preprint arXiv:1909.01259*.

[14] Attardi, G., Buzzelli, A., & Sartiano, D. (2013). Machine Translation for Entity Recognition across Languages in Biomedical Documents. In *CLEF (Working Notes)*.

[15] Olszewski, Robert. (2001). Generalized feature extraction for structural pattern recognition in time-series data.

[16] PolicyCLOUD. D3.6 – PolicyCLOUD’s Societal and Ethical Requirements & Guidelines – M22, Audino Alice, 2021.

[17] European Union Agency for Cybersecurity, *EUCS – Cloud Services Scheme, https://www.enisa.europa.eu/publications/eucs-cloud-service-scheme*, retrieved 2022-06-09.

[18] PolicyCLOUD. D4.3 – Reusable Model & Analytical Tools: Design and Open Specification 2, Biran Ofer, 2021.

[19] PolicyCLOUD. D8.1 – POPD-Requirement No. 1, Meerkamp Marc, 2021.

[20] PolicyCLOUD. D2.2 – Conceptual Model & Reference Architecture. Argyro Mavrogiorgou, Panayiotis Tsanakas, Panayiotis Michael, Vrettos Moulos, et. al. 2020.

[21] PolicyCLOUD. D2.6 – Conceptual Model & Reference Architecture. Argyro Mavrogiorgou, Panayiotis Tsanakas, Panayiotis Michael, Vrettos Moulos, et. al. 2021.

[22] PolicyCLOUD. D3.4 – Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 2, Ledakis Giannis et al. 2021.

[23] PolicyCLOUD. D2.5 – State of the art & Requirements Analysis, LXS, 2021.

[24] PolicyCLOUD. D6.11 – Use Case Scenarios Definition & Design, Sancho Javier et al. 2022.

[25] PolicyCLOUD. D6.10 – Use Case Scenarios Definition & Design, Sancho Javier et al. 2021.