Policy Cloud

Cloud for Data-Driven Policy Management

CLOUD FOR DATA-DRIVEN POLICY MANAGEMENT

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

D2.2 CONCEPTUAL MODEL & REFERENCE ARCHITECTURE

Dissemination Level	PU
Due Date of Deliverable	31/8/2020, M08
Actual Submission Date	4/9/2020
Work Package	WP2 Requirements, Architecture & Innovation
Task	T2.2 Definition of Target Conceptual Model & Reference Architecture
Type	Report
Approval Status	
Version	V1.1
Number of Pages	p.1 – p.53

Abstract: This document provides the Conceptual Model and Reference Architecture of PolicyCLOUD (Deliverable D2.2). More specifically, the document provides the definition of the overall architecture and its components and presents example scenarios and associated data sources from the project's Use Cases, that will be used for end-to-end data path analysis in next releases. Updates of the deliverable will be published in M18 and M30.

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

Versioning and Contribution History

Version	Date	Reason	Author
0.5	21/01/2020	Initial Architecture, PolicyCLOUD Kick-Off Meeting	Thanos Kiourtis (UPRC), Argyro Mavrogiorgou (UPRC), George Manias (UPRC), Dimosthenis Kyriazis (UPRC)
0.6	30/03/2020	Cloud based environment	Giuseppe La Rocca (EGI)
0.7	10/04/2020	Constituent architecture - Data Acquisition and Analytics Layer	Ofer Biran (IBM)
0.7	10/04/2020	Incentives Management, Situational Knowledge Analysis, Opinion Mining, Sentiment Analysis	Maria Angeles Sanguino Gonzalez (ATOS), Jorge Montero Gomez (ATOS), Tomas Pariente Lobo (ATOS)
0.7	10/04/2020	Stakeholders Data Stores (ARAGON)	Rafael del Hoyo (ITA)
0.7	10/04/2020	Data Governance Model, Protection and Privacy Enforcement	Konstantinos Theodosiou (UBI), Giannis Ledakis (UBI)
0.7	10/4/2020	DataStores, DataStores integration	Javier López Moratalla (LXS), Sadra Ebro (LXS), Patricio Martinez (LXS)
0.7	10/04/2020	Policy Development Toolkit and Data Visualization	Konstantinos Moutselos (ICCS), Ilias Maglogiannis (UPRC)
0.7	10/04/2020	Cloud Gateways, Sources Reliability, Data Cleaning, Data Interoperability, Social Dynamics	Thanos Kiourtis (UPRC), Argyro Mavrogiorgou (UPRC), George Manias (UPRC), Dimosthenis Kyriazis (UPRC)
0.7	10/04/2020	Framework for Cloud Use by Public Authorities	Petya Bozhkova (OKS)
0.7	10/04/2020	Modelling & Design of Middleware for Policies	Konstantinos Nasias (OKS)
0.7	14/04/2020	Technical Coordination, Architecture Integration, Editing of Internal Document	Panayiotis Tsanakas (ICCS), Panayiotis Michael (ICCS), Vrettos Moulos (ICCS)
Note: Version 0.7 of Deliverable D2.2 was prepared and delivered as an internal report to all partners in M03 as per the Grant Agreement for PolicyCLOUD.			
1.0	07/08/2020	Initial Architecture, PolicyCLOUD Kick-Off Meeting	Thanos Kiourtis (UPRC), Argyro Mavrogiorgou (UPRC), George Manias (UPRC), Dimosthenis Kyriazis (UPRC)
1.0	07/08/2020	Cloud based environment	Giuseppe La Rocca (EGI)
1.0	07/08/2020	Constituent architecture - Data Acquisition and Analytics Layer	Ofer Biran (IBM), Oshrit Feder (IBM)
1.0	07/08/2020	Incentives Management, Situational Knowledge Analysis, Opinion Mining, Sentiment Analysis	Maria Angeles Sanguino Gonzalez (ATOS), Jorge Montero Gomez (ATOS),

Version	Date	Reason	Author
			Tomas Pariente Lobo (ATOS)
1.0	07/08/2020	Data Governance Model, Protection and Privacy Enforcement	Konstantinos Theodosiou (UBI), Giannis Ledakis (UBI)
1.0	07/08/2020	DataStores, DataStores integration	Javier López Moratalla (LXS), Sadra Ebro (LXS), Patricio Martinez (LXS)
1.0	07/08/2020	Policy Development Toolkit and Data Visualization	Konstantinos Moutselos (ICCS), Ilias Maglogiannis (UPRC)
1.0	07/08/2020	Cloud Gateways, Sources Reliability, Data Cleaning, Data Interoperability, Social Dynamics	Thanos Kiourtis (UPRC), Argyro Mavrogiorgou (UPRC), George Manias (UPRC), Dimosthenis Kyriazis (UPRC)
1.0	07/08/2020	Framework for Cloud Use by Public Authorities	Petya Bozhkova (OKS)
1.0	07/08/2020	Modelling & Design of Middleware for Policies	Konstantinos Nasias (OKS)
1.0	07/08/2020	Use Case Scenarios Description - Participatory Policies Against Radicalization	Armend Duzha (MAG), Nikos Achilleopoulos (MAG)
1.0	07/08/2020	Use Case Scenarios Description - Intelligent Policies for The Denomination of Origin	Rafael del Hoyo (ITA), Javier Sancho (SARGA)
1.0	07/08/2020	Use Case Scenarios Description - Urban Policy Making Through Analysis of Crowdsourced Data	Iskra Yovkova (SOF), Petya Bozhkova (OKS)
1.0	07/08/2020	Use Case Scenarios Description –Predictive Analysis Towards Unemployment Risks Identification and Policy Making	Ebenezeer Williams (LON), Sarah Frost (LON), Adil Mohammed Ali (LON)
1.0	07/08/2020	Technical Coordination, Architecture Integration, Editing of Internal Document	Panayiotis Tsanakas (ICCS), Panayiotis Michael (ICCS), Vrettos Moulos (ICCS)
1.1	04/09/2020	Quality Check performed. Changes addressed.	Argyro Mavrogiorgou (UPRC), Panayiotis Tsanakas (ICCS), Panayiotis Michael (ICCS), Vrettos Moulos (ICCS)

Author List

Organisation	Name
UPRC	Thanos Kiourtis
UPRC	Argyro Mavrogiorgou
UPRC	George Manias
UPRC	Dimosthenis Kyriazis

Organisation	Name
EGI	Giuseppe La Rocca
IBM	Ofer Biran
IBM	Oshrit Feder
ATOS	Maria Angeles Sanguino Gonzalez
ATOS	Jorge Montero Gomez
ATOS	Tomas Pariente Lobo
MAG	Armend Duzha
MAG	Nikos Achilleopoulos
UBI	Konstantinos Theodosiou
UBI	Giannis Ledakis
LXS	Javier López Moratalla
LXS	Sadra Ebro
LXS	Patricio Martinez
UPRC	Konstantinos Moutselos
UPRC	Ilias Maglogiannis
ITA	Rafael del Hoyo
SARGA	Javier Sancho
OKS	Petya Bozhkova
OKS	Konstantinos Nasias
SOF	Iskra Yovkova
LON	Ebenezeer Williams
LON	Sarah Frost
LON	Adil Mohammed Ali
ICCS	Panayiotis Tsanakas
ICCS	Panayiotis Michael
ICCS	Vrettos Moulos

Abbreviations and Acronyms

Abbreviation/Acronym	Definition
ABAC	Attribute-based access control
API	Application Programming Interface
CMF	Cloud Management Framework
DB	Database
DSS	Decision Support System
EC	European Commission
EOSC	European Open Science Cloud
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
PM	Policy Model
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

SQL	Structured Query Language
TRL	Technology Readiness Level
VM	Virtual Machine

Contents

List of Tables

List of Figures

1 Executive Summary

The PolicyCLOUD Conceptual Model & Reference Architecture are presented in this report (Deliverable D2.2) released in M8 of the project. Updates of the deliverable will be published in M18 and M30.

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 model the 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 is exploited for the definition of forward-looking policies which are dynamically adapted and methodically focused on the population that 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 includes a Data Governance Model, Protection and Privacy Enforcement and the Ethical Framework.

The architecture allows for integrated acquisition and analytics, as also data fusion with processing and initial analytics combined with seamless analytics on hybrid data at rest.

A number of scenarios developed for the Use Cases of PolicyCLOUD are also included in the document, in order to serve as end-to-end examples for the demonstration of the data ingest flow and data exploitation and for the analysis of the processing and data transformations along the complete data path. Problem statement, main objectives, Key Performance Indicators and data sources sections are provided for each scenario in this first release of Deliverable 2.2 while the detailed end-to-end data path analysis will be prepared during the completion of the first prototypes in M10 and will be included in the next update of the document in M18.

The overall architecture and Use Case scenarios have been discussed by all partners (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 internal workshops integrating constituent architectures and specialized internal workshops discussing Architecture integration and end-to-end Use Case journey. Finally, conclusions are provided in the last section of the document.

2 Introduction

This document is the first published report of the Conceptual Model and Reference Architecture of PolicyCLOUD (Deliverable D2.2) which is based on the preliminary specification of the PolicyCLOUD architecture made internally available to partners in M3. 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. Updates of the deliverable will be published in M18 and M30.

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**-Policies 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 Framework** presented in Section 7.5 is included in the architecture from the very beginning in order to provide end-to-end ethical considerations and legal know-how to the project.

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 acquisition and analytics, as also data fusion with processing and initial analytics combined with seamless analytics on hybrid data at rest. A number of **scenarios developed for the Use Cases of PolicyCLOUD** are presented in section 8, in order to serve as end-to-end examples for the demonstration of the data ingest flow and data exploitation and for the analysis of the processing and data transformations along the complete data path. These scenarios have been prepared during specialized internal workshops discussing Architecture integration and end-to-end Use Case journey. At this first release of Deliverable D2.2 Use Case scenarios are focused on problem statement, main objectives, Key Performance Indicators and data sources to be used serving as a basis for a detailed end-to-end data path analysis that will be prepared during the completion of the first prototypes in M10 and will be included in the next update of the deliverable in M18. Finally, in Section 9 the conclusions of this report are presented.

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 [35].

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, obtain health analytics outcomes and perform policy modelling and policy making.

Object Storage is designed to support exponential data growth and cloud-native workloads. It provides built-in high-speed file transfer capabilities, 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** [36].

4 PolicyCLOUD offerings

PolicyCLOUD offerings are materialized through five main building blocks, supported by the **Ethical 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.

Finally, the **Ethical Framework** assures that all PolicyCLOUD offerings conform to the required ethical, legal and security aspects 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** tasks integrated with initial analytics and data processing tasks.
- **Seamless Analytics** on both 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.
- **Framework for Cloud use 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.
- **Modelling and Design of Middleware for 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 the System. 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 Framework and a Data Governance Model providing all above capabilities.

6 PolicyCLOUD Conceptual Model

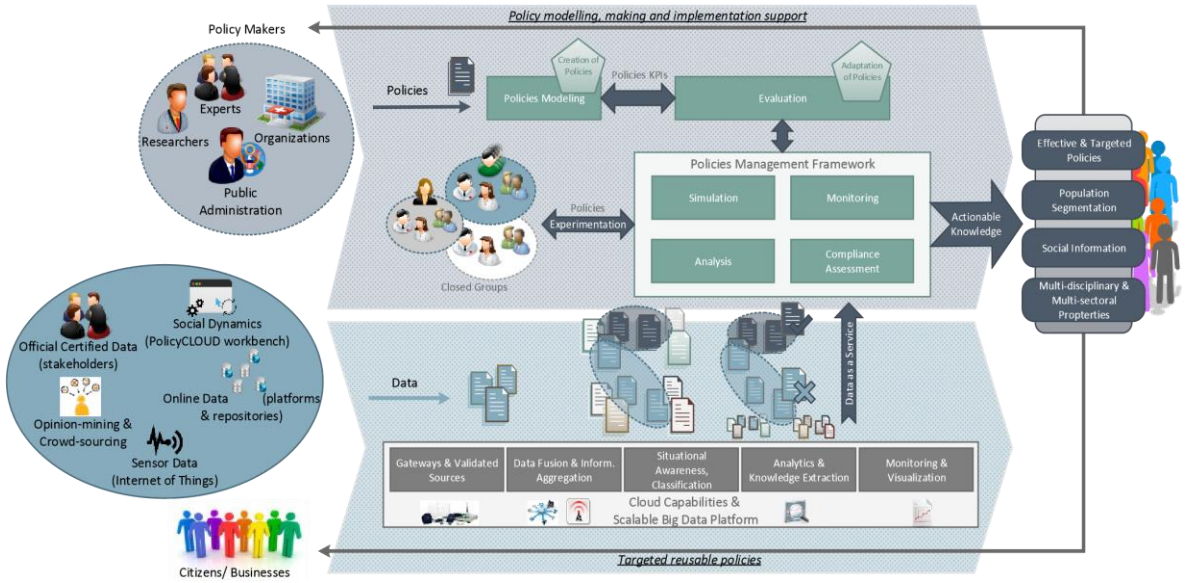

FIGURE 1 – THE POLICYCLOUD CONCEPTUAL MODEL

6.1 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 any source.

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 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 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). Data clearance processing combines legal 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 with respect to data and to the state of existing policies that correspond to target groups / public authorities with specific goals and population characteristics.

Thus, data processing exploits policies collections / clusters. 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, analysed and evaluated.

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 PolicyCLOUD 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 fine-grained to create variations with different parameter sets, which will be 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).

7 PolicyCLOUD Architecture

7.1 Architecture Building Blocks

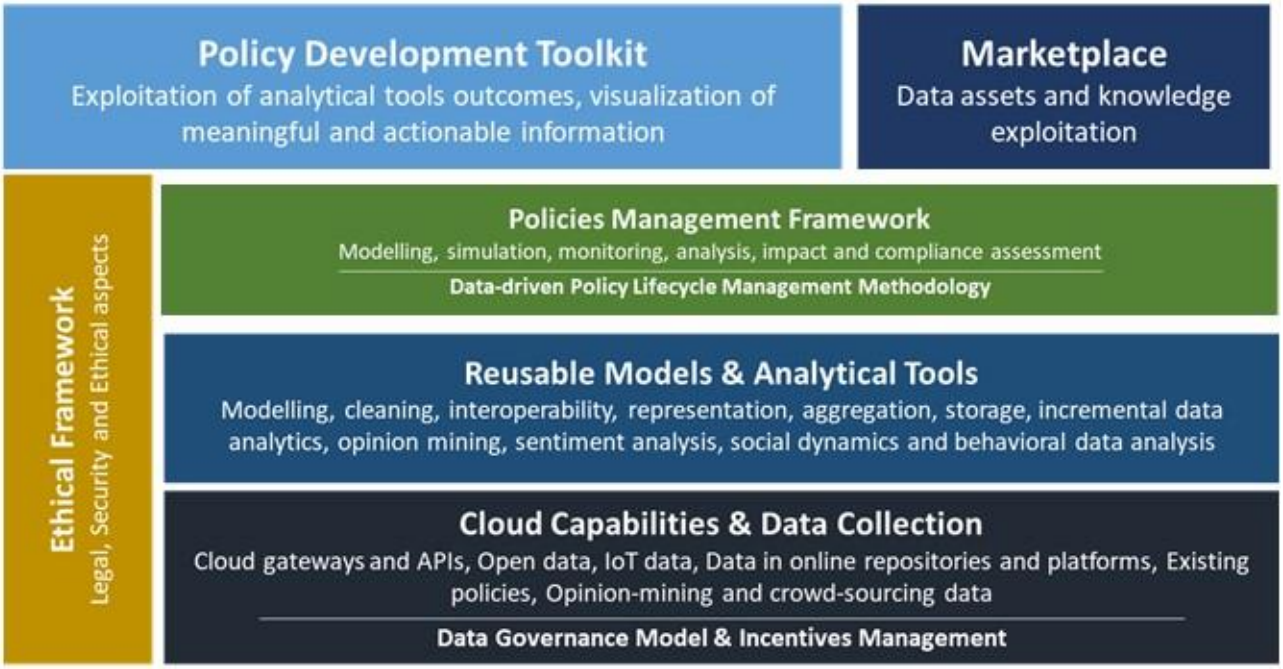

FIGURE 2 – POLICYCLOUD ARCHITECTURE BUILDING BLOCKS

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 (e.g. local, national,

etc). 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, and citizens to interact with the models and analytical tools as well as to specify their requirements and constraints with respect to different policies (e.g. specification of the need for policies that can have a real-time impact due to emergencies). 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.

The **Ethical Framework** assures that all the PolicyCLOUD offerings conform to the required ethical, legal and security aspects, thus ensuring the sustainability of the modelled policies. The framework is **vertically** depicted in the figure given that it obtains information from the **Cloud Capabilities & Data Collection Engine** (such as social networks data), while it communicates in a **bi-directional way with the Interacting Environment** by obtaining data from the **Policy Development Toolkit** and the **data marketplace**, and by specifying analytics tasks through this toolkit.

The architecture building blocks will be implemented over the European Cloud Initiative infrastructure offered by EGI (Figure 3).

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.

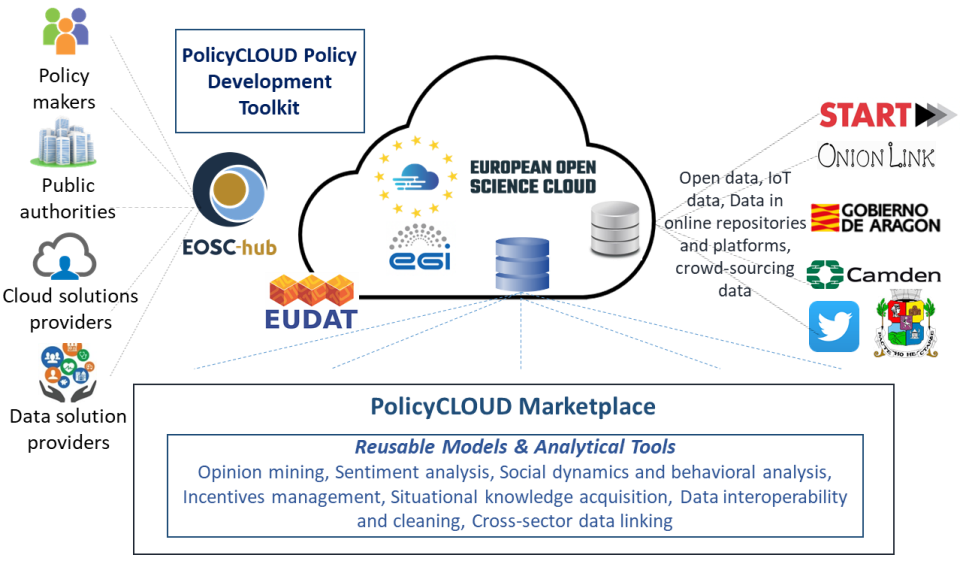

FIGURE 3 – POLICYCLOUD ARCHITECTURE IMPLEMENTATION OVER THE EUROPEAN CLOUD INITIATIVE INFRASTRUCTURE OFFERED BY EGI

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. Updates of the architecture will be published in M18 and M30. The architecture's layers and frameworks will be analyzed in the sections that follow.

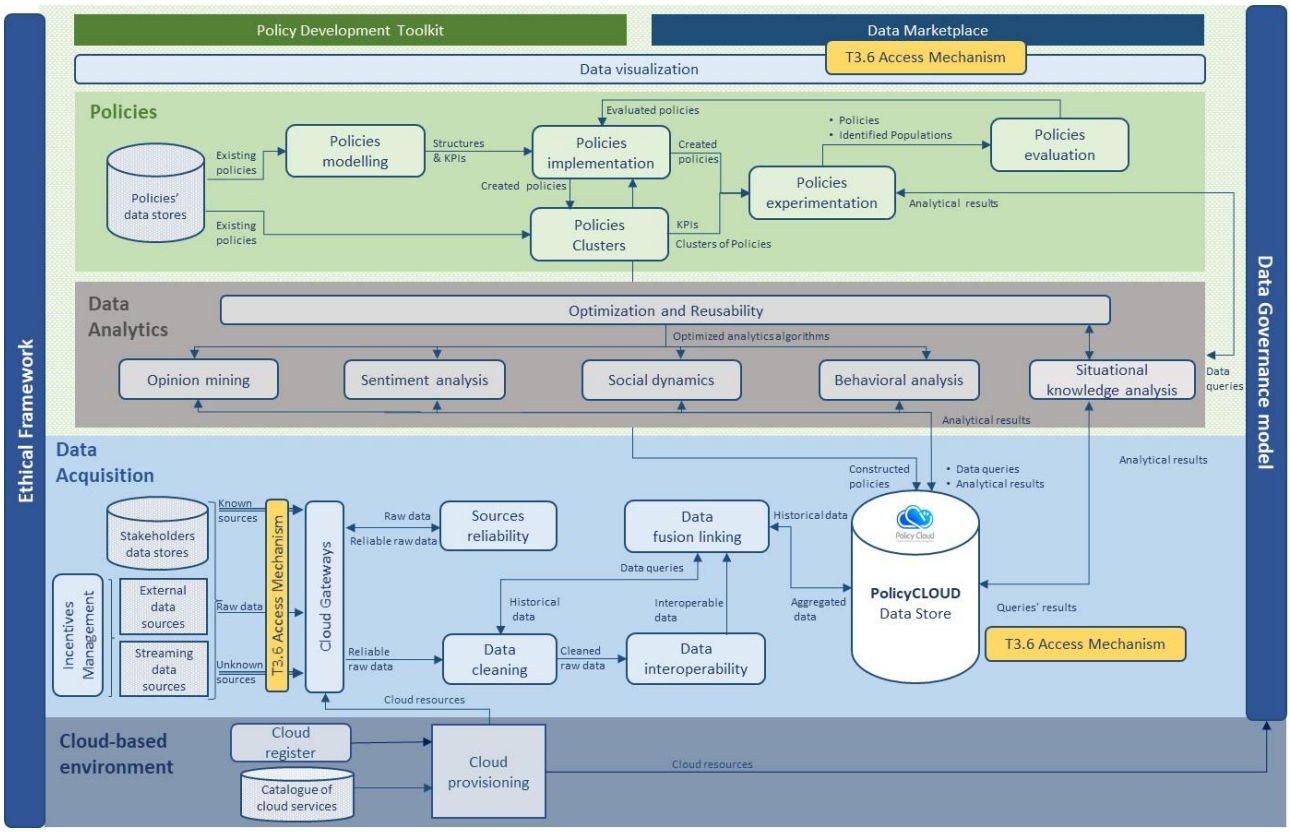

FIGURE 4 – POLICYCLOUD OVERALL ARCHITECTURE

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 (e.g. mobile devices, IoT sensors, etc.) 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 exploiting the *sources reliability* component 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 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* that provides the respective services obtained from a catalogue of cloud infrastructure resources. 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 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* 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.

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 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 end-points.

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 is contributing to the provisioning of the needed computing resources to set-up the PolicyCLOUD infrastructure. This cloud infrastructure will help policy makers, 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 will be highlighted in D3.1 - Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 1.

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

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

The Cloud Gateway and API component developed will 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. It will act as the only entryway into PolicyCLOUD project allowing multiple APIs or microservices to act cohesively and provide a uniform, gratifying experience to each stakeholder. The provided Gateway API will allow 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 will enhance the design of resources and structure, add dynamic routing parameters and develop custom authorizations logic. PolicyCLOUD's Cloud Gateway and API component will support scalability, high availability and shared state without compromising performance. Moreover, it will support 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 [37], a framework that bases its functionality on microservices architecture methodology, will be 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 or fault tolerance [37].

Through this ability the component will, also, be able to directly ingest incoming data into the appropriate data store based on their privacy level. 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, this component will examine and capture 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. To this context, the Gateway will be 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 will be 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, as well as the collection of the data that comes only from reliable data sources so as to be used for further analysis.

7.4.2 Incentives Management

The overall idea of Incentives Management is to offer a set of tools to identify and manage incentives able 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 analyze participation behaviors (interest and sentiment), aiming at involving different stakeholders in the policies and mechanisms to declare incentives and manage them. This means that the task will investigate and work on a set of initiatives and services towards that end, such as:

- Activities for the engagement of citizens.
- Tools for identifying trends and tracking events.
- Mechanisms to allow the participants to declare their incentives.
- A tool to manage incentives to make proposals on a per-sector, per-participant or per-population basis.

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 [11] (Malone, 2010): what, who, why, and how, the incentives plan will be pre-established as follow:

- **WHO (participants/requesters):** The Incentives Management task will be focused in 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. For the PolicyCLOUD project two actions are required from the participants.
    - evaluate existing policies
    - suggest requirements for new ones
- **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".* Figure 5 shows the big picture of this concept.

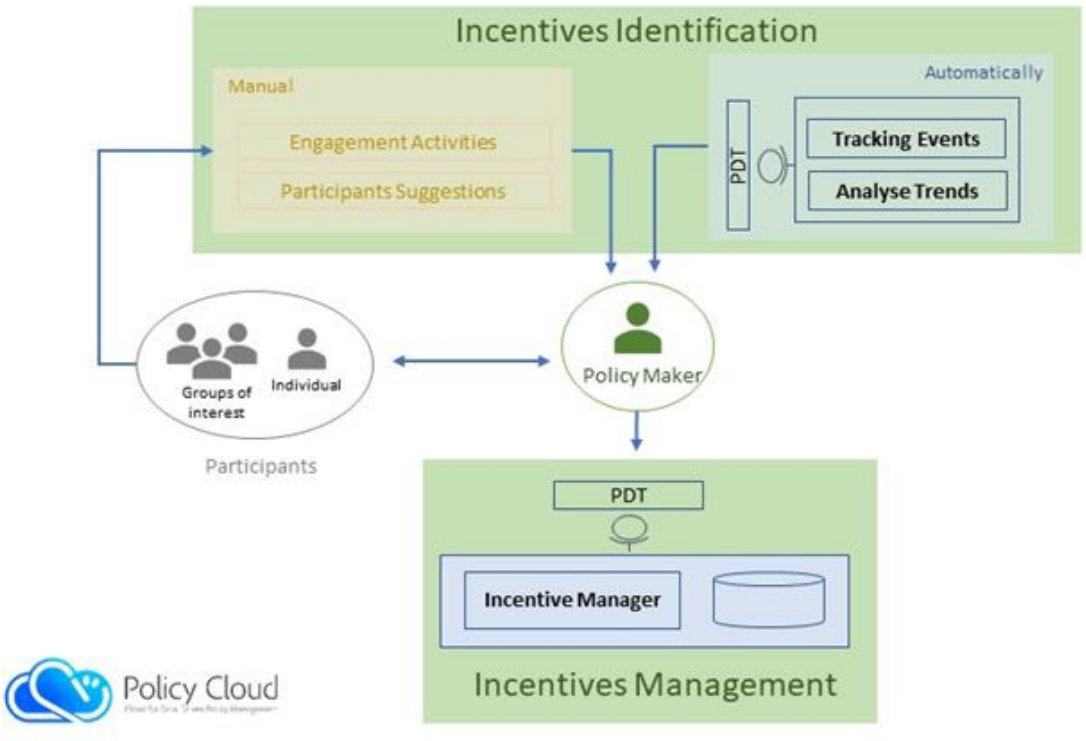

FIGURE 5 – INCENTIVES IDENTIFICATION AND MANAGEMENT BIG PICTURE

For more details, please refer to Deliverable D3.1 "Cloud Infrastructure Incentives Management and Data Governance: Design and Open Specification 1".

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 by different and heterogeneous manners. Firstly, as part of the project itself, four different use case providers are planned to be integrated to the common platform, while on the same time, the platform itself is envisioned to be exploited in the future by other cases. Each one of those independent organizations 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. Streaming data refers also to external data that is remotely accessed upon demand. Nevertheless, 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, typically the data are being transferred to a data warehouse such as Object Storage, that is 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 and must remain on premise due to data regulator constraints. The central data store component needs 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. 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 property of 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 some 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 will overcome the need for accessing data that are stored in different silos inside an organization or in external sources. The polyglot capabilities of the data store are also important for the datalake capability of enabling query processing of unstructured data, which is typically used in the majority of the datasets provided by the existing and future scenarios.

- ***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, where usually the solution of caching the results is preferred over the consistency of the result with respect to transactional semantics. This is very important when pre-processing needs to take place over a stream, which is a typical scenario that PolicyCLOUD targets.

Apart from the dealing with those three different types of data, the data management mechanisms of the platform will benefit from the results for the EU H2020 project named BigDataStack and its Seamless Analytical Framework, where similar scenarios with regards to the movement of obsolete data from an operational to an object store are being addressed. That will allow for data to be moved to the object storage on 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 will 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 will be further developed to cope with the scenarios defined here, with the plan to increase its current technology readiness level (TRL).

7.5 Ethical Framework

7.5.1 Ethical Framework

The **Ethical and legal compliance framework** (task T3.5) is aiming at analysing and giving guidance on the legal, ethical and societal requirements with regard to the infrastructure. Particular attention is currently being paid to the choice of data and the sources it originally comes from, as well as the admissibility of its use by the controllers/contributors bringing in data into the PolicyCLOUD infrastructure.

This counts for personal data in terms of the General Data Protection Regulation ("GDPR"), as well as for the use of "open data" that involves legal issues with regard to the protection of intellectual property including the protection of databases and trade secrets. At this phase of the development of the architecture the processing of personal data in terms of the GDPR is currently out of scope. However, it appears that a clearing mechanism will have to be implemented for private data, as well as for "open data". Task T3.5 will also consider – with the help of partners – guidance on the evaluation if and to which extent personal data in terms of the GDPR will or will not be processed.

The clearing mechanism for all types of data could be implemented at the stage of "data acquisition". It will have to be determined if this takes place before or after the data is being processed through the cloud gateway. Besides, task T3.5 will aim at lining out the (shared) responsibilities of PolicyCLOUD, the partners and stakeholders when processing personal data in terms of the GDPR, as well as basic considerations for the admissibility of use of such data. This will also include the role of the cloud provider as a processor. Task T3.5 will also assess – together with the other partners - the options of an anonymization of data as well as the position of such mechanism within the PolicyCLOUD infrastructure in order to avoid the (partial) application of the GDPR within the infrastructure. Moreover, guidance will be given on the requirements of security and confidentiality, as well as data protection by design and default, taking into account the work in task T3.6 (Data Governance Model, Protection and Privacy Enforcement).

Furthermore, task T3.5 will analyse other ethical as well as societal requirements in order to make the PolicyCLOUD infrastructure compliant with these and maximizing societal acceptability and trust in PolicyCLOUD and through it in policies. This will inter alia include issues on the reliability of data and prevention of false raw data in order to mitigate the risk of incorrect policies or factual bases of a decision based on the policy. Attention will also have to be paid to the abuse of data in order to intentionally manipulate a decision- making process. Besides, the risk of an abuse of the platform as a whole for unethical issues including the abuse by policy makers will have to be determined. Further analysis will be necessary with regard to the data marketplace as well as the liability for data in a later period of commercial exploitation.

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).

An extract from the overall architecture is shown in Figure 6 for convenience. The extract 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 7 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.

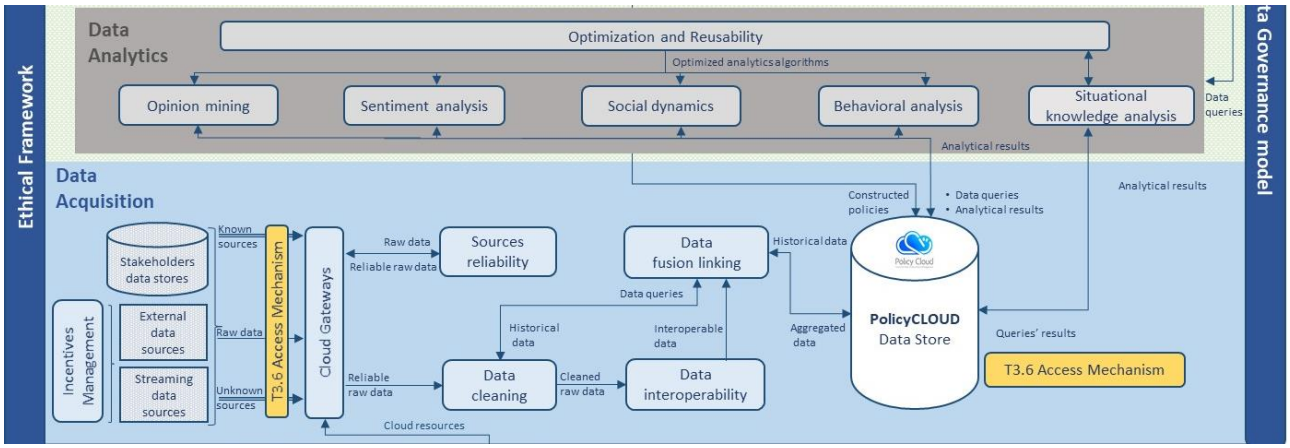

FIGURE 6 – EXTRACT FROM THE POLICYCLOUD OVERALL ARCHITECTURE DIAGRAM

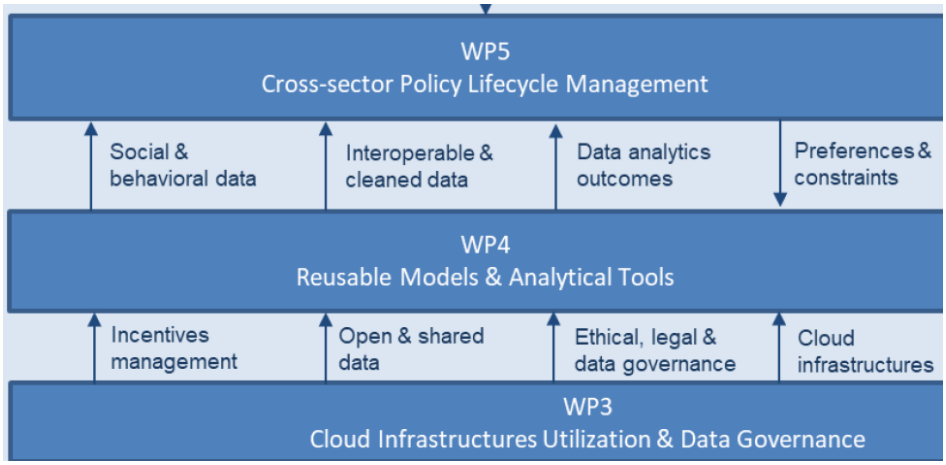

FIGURE 7 – WP4 INTERFACE WITH WP3 AND WP5

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.

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. Rest data analytic function 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 will offer all the appropriate algorithms and techniques for detecting and correcting (or removing) corrupt or inaccurate records from all the collected data that will be retrieved as an input from the Cloud gateways component. More specifically, this component will be 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 will be 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 will detect and totally clean 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 will be automatically detected and erased. Finally, since any possible inconsistent data will be 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 it will be 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 [47], 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 [48], [49] 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

In this section we present 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 (https://spark.apache.org/streaming), KSQL (https://github.com/confluentinc/ksql) or Serverless engine based on Apache OpenWhisk (https://openwhisk.apache.org). In Figure 8 we depict the end-to-end data path for this scenario.

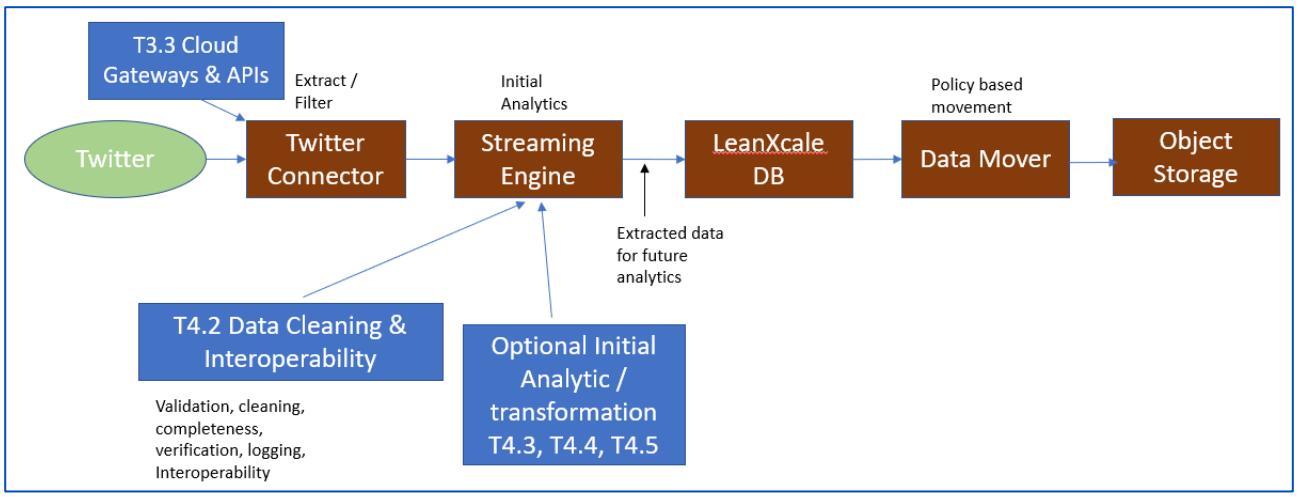

FIGURE 8 – THE STREAMING DATA PATH

Task 4.1 (Cross-sector Data Fusion Linking) is responsible for the overall data path and streaming framework in this scenario. The Twitter connector will be implemented by task T3.3 (Cloud Gateways) and will create 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 will be performed by Task T4.2 which will be running within the streaming engine. Optional initial analytics on the streamed data may be performed by tasks T4.3 (Situational Knowledge Acquisition & Analysis), T4.4 (Opinion Mining & Sentiment Analysis) and T4.5 (Social Dynamics & Behavioural Data Analytics).

At the end of the data path, the Data Mover is responsible for moving older data from hotter storage (DB) to a colder (object storage) 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 still after cleaning and validation processes). The "right" side of the data path in Figure 9 present a periodical movement of older data from hotter storage (DB) to a colder (object storage) according to policy rules, which addresses 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 (https://bigdatastack.eu) and will be 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 9 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.

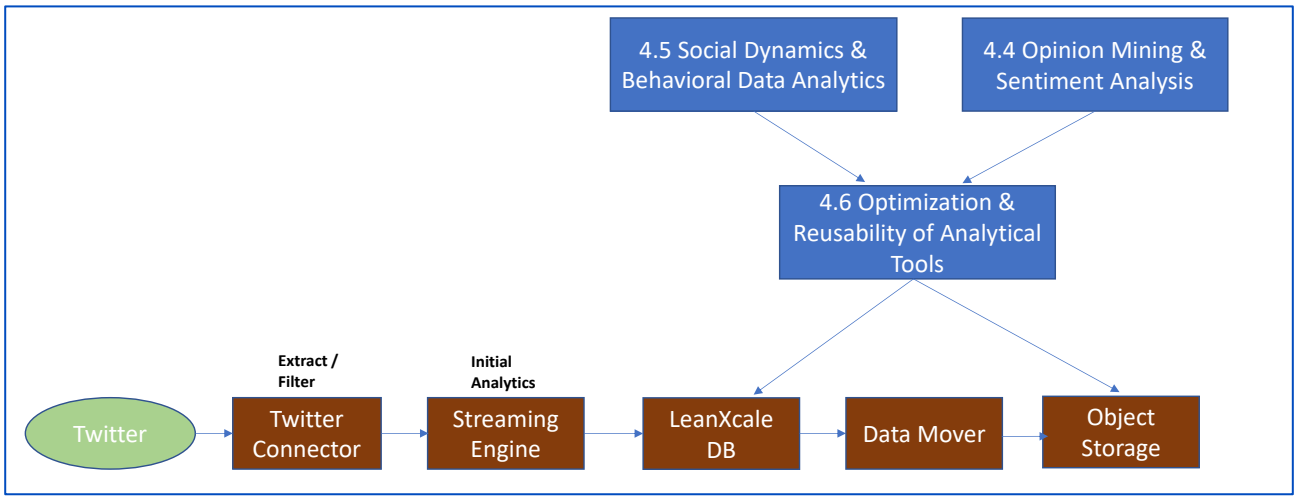

FIGURE 9 – SEAMLESS ANALYTICS ON INGESTED DATA

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 will be used to influence the decisions taking place based on the PolicyCLOUD system.

The following capabilities will be provided through the SKA component:

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

Some of the characteristics of the component will be:

- Feature Extraction. The extraction knowledge stage will be done through Feature Extraction (ML) techniques able to create/derive new situational features from existing ones. This extraction step will be enhanced by the situational knowledge model which will 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 [4] (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), in order 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 will perform a sentiment analysis based on the input received from the pilots about their policies. This input could come from what the citizens say in social media channels, from platforms owned by the pilot (getting feedback on various subjects), or other channels that will be 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 will be 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 will take 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 will have 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 will be stored in the database provided by PolicyCLOUD to be used by other components.

7.6.10 Social Dynamics

The Social Dynamics component will consist of a concurrent, web-based environment for social simulation. The environment will allow the user to create graph-based population models online. These models will 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 will be able to upload appropriately structured population data from databases conforming with these parameters. Individual characteristics will consist of sets of variables that capture the relevant attributes for each individual in the model. Link characteristics will specify a set of variables used for the creation of weighted links between individuals. Individual dynamics will 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 will 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 will allow users to specify all these parameters online in the simulation environment. Based on these specifications, the environment will be 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 will exploit opportunities for the breakdown of the tasks in each simulation into concurrent units that will allow the simulator to optimize its use of computational resources.

7.7 Layer 3 – Policies 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 Framework for Cloud Use by Public Authorities

Within the framework, a view of the current cloud infrastructures and big data technologies used from public administrations will be developed 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.

The overall architecture will evaluate how the value of this Framework and associated report will be exploited.

7.7.2 Modelling & Design of Middleware for Policies

A middleware based on .NET Core will be created as the adapter pattern, to retrieve data from the Policies data stores. 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. This will allow the end users to create a policy by selecting a schema of data, applying well known Key Performance Indicators or set some with simple linear functions and create a set of rules (criteria). As for the existing policies the users will name a description with a set of rules (criteria) which will apply the values of a specific schema of data and Key Performance Indicators. This will provide output to be used by Task 5.3.

For more details, please refer to Deliverable D5.2 "Cross-sector Policy Lifecycle Management: Design and Open Specification 1".

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) constitutes the Front-End of the System. It is the component that integrates several sub-components to enable policy makers (PMs) to create, update and validate policy models. As a Decision Support System (DSS) for evidence-based public decision-making, 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 be able to set Key Performance Indicators (KPIs) that support the policy in focus. KPIs 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.

For the visualization of analytical tools outcomes, the PDT will provide a reporting tool that will enable to build visual analytical reports. The reporting will be produced from analytical queries and will include summary tables as well as graphical charts resulted from the Analytics. The policy monitoring dashboard will be adaptable, since it will enable to build a custom dashboard that can include 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 Layer, the Datastore Backend (LXS) 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 is the frontend of the PolicyCLOUD platform intended to be used mostly by the policymakers, who are the main target population of the project. Policymakers will be able to create and evaluate Policy Models (PMs), while also keep a history of their transactions. PDT will hide the complexity of the system dataflow to provide to the policymaker/user of the PolicyCLOUD platform 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 10. PDT may be considered as the point of integration and interaction of the platform with the policymakers. Through the PDT, the policymakers will be able to question the platform data and exploit the analytics tools to perform policy creation and evaluation.

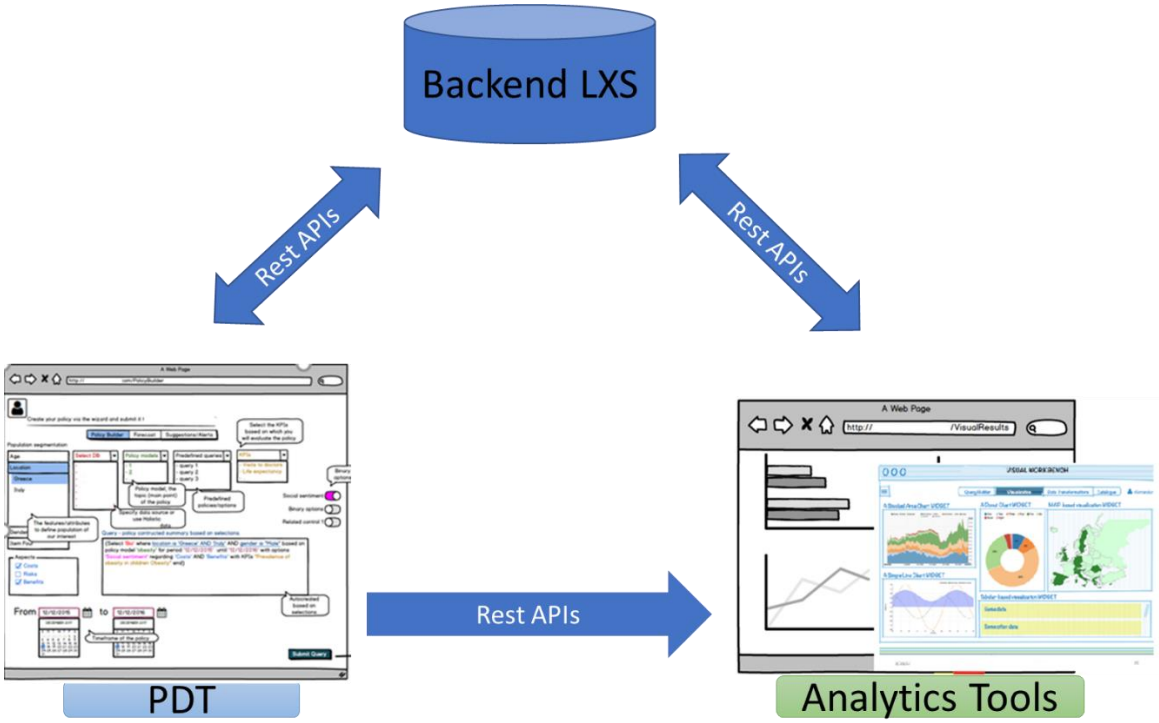

FIGURE 10 – POLICY DEVELOPMENT TOOLKIT COMMUNICATION COMPONENTS

Figure 10 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 selected data source with the parameters specified by the policymaker. The arrows in Figure 10 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 WP4.

The Policies will be serialized in a predefined format following common syntax (in JSON) into the Datastore. The PDT will translate/deserialize the policy objects retrieved from the Datastore into UI objects to provide the visual environment for the policymaker actions.

The arrow between PDT and LXS 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 [10] 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.

7.9 Layer 5 - Data Marketplace

Components: Data Marketplace (T7.2)

7.9.1 Data Marketplace

The Data Marketplace component will consist of a concurrent, web-based environment for providing functions and interfaces in order to support searching, retrieving and ingesting of datasets and policies. Since PolicyCLOUD aims at a cloud-based data-driven policy management, it enables the stakeholders to model, analyze, evaluate and optimize their policies using a variety of Big Data tools and services. Hence, the Data Marketplace is being adapted to support the storage and retrieval of valuable data artefacts and to enable data and knowledge to be exploited as assets. Moreover, the Data Marketplace will be aligned with the principle of persistent storage by extending the metadata and semantics that are being attached to the datasets and policies with valuable fields for successful data integration, accuracy in the format of the document and policy, identified KPIs, etc. Thus, it enables the development of enhanced operations for Creating, Retrieving, Updating and Deleting (CRUD) metadata inside it, while re-assures the correct data format of the stored datasets and policies. Furthermore, since the types of data and policies vary, one of the requirements, satisfied by the Data Marketplace, will be the full-text search capabilities in structure-agnostic assets. Hence, it will provide seamless retrieval abilities in deep-hierarchical machine-readable document structures and thus interoperability of datasets and policies will be enhanced. Finally, the Data Marketplace is responsible to store different objects in regard to policies management across the whole lifecycle.

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 11 (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 11 (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 11 (B) - is the access control being also provided on the level of data visualization, thus allowing or denying access to specific data analytics. The third - Figure 11 (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 (private data store is not included in this first release of the architecture).

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.

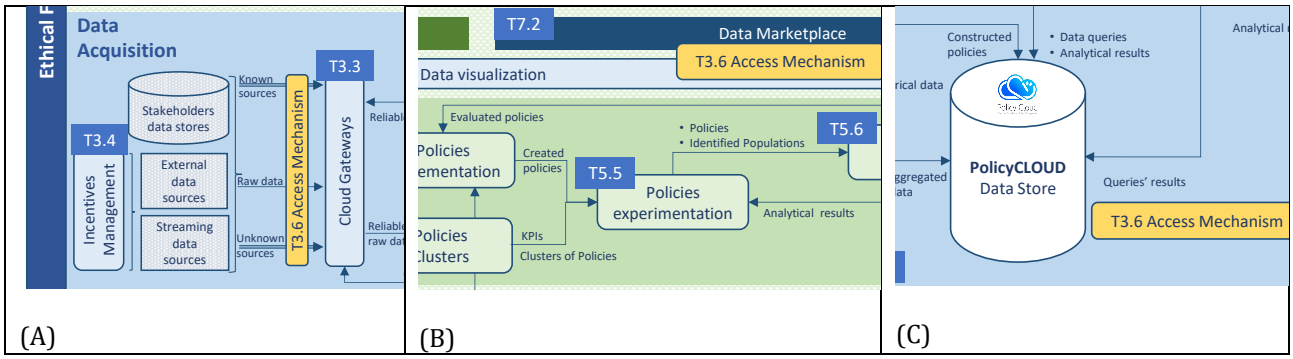

FIGURE 11 – DATA GOVERNANCE MODEL, PROTECTION AND PRIVACY ENFORCEMENT MECHANISMS – EXTRACTED VIEWS (A), (B) AND (C) FROM THE DIAGRAM OF POLICYCLOUD OVERALL ARCHITECTURE.

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

A number of scenarios developed for the Use Cases of PolicyCLOUD will be presented, in order to serve as end-to-end examples, demonstrating the data ingest flow and data exploitation while analysing the processing and data transformations along the complete data path.

At this first iteration of the project, as presented in the following sections, the Use Case scenarios are focused on problem statement, main objectives, Key Performance Indicators and data sources to be used, while a detailed end-to-end data path analysis will be prepared during the completion of the first prototypes in M10 and will be included in the next update of Deliverable D2.2 in M18.

8.1 Participatory Policies Against Radicalization

8.1.1 Scenario 1.1 - Problem Statement

Need for a Heatmap visualization, that maps the frequency of occurrence of incidents targeting vulnerable groups (e.g. children, minor) in the geographic proximity of a town/region.

8.1.2 Main Objective

Validate existing (local/regional/national and EU) policies to counter violent-extremism and investigate if there is a need to adjust / update them or create new ones[3] based on the information **extracted from open data**[4] e.g. the Global Terrorism Database (GTD).

8.1.3 Key Performance Indicators

KPI 1.1.1	Number of occurrences of incidents that can affect children in an area[5] > 0
KPI 1.1.2	Number of active groups / initiatives in an area > 0

[3] Examples of new policies could be the promotion of new training programmes / social activities for children to be adopted by schools.
[4] In this first scenario we are considering only the data coming from the GDT, as they are structured and trusted. As a second scenario we can consider the information retrieved from social media channel and as a third one a mix of both cases (to be investigated how critical is this to be implemented).
[5] The area of interest shall be defined by the policy maker. The system may propose / suggest some alternatives (e.g. within the city, region or specific districts such as sub-urban areas).

8.1.4 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Participatory Policies Against Radicalization	Scenario #1.1	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 1 – DATA SOURCES LIST FOR SCENARIO 1.1 OF THE PARTICIPATORY POLICIES AGAINST RADICALIZATION USE CASE

8.1.5 Scenario 1.2 – Problem Statement

Identify radicalization efforts / incidents targeting vulnerable groups (e.g. children, minor) and the main parties (e.g. individuals and groups) involved.

8.1.5.1 Main Objective

Validate existing (local/regional/national and EU) policies to counter violent-extremism and investigate if there is a need to adjust / update them or create new ones based on the assessment of ***social media channels (e.g. Twitter, Reddit, etc.)*** observations against perceived radicalization efforts.

8.1.5.2 Key Performance Indicators

KPI 1.2.1	Number of identified users and groups / initiatives in Europe > 0
KPI 1.2.2	Number of new terms / keywords / hashtags identified from the policy maker > 0
KPI 1.2.3	Performance of sentiment analysis / opinion mining on comments: at least weakly

8.1.6 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Participatory Policies Against Radicalization	Scenario #1.2	Data from Twitter RSS Feeds	Data from Twitter RSS Feeds

TABLE 2 – DATA SOURCES LIST FOR SCENARIO 1.2 OF THE PARTICIPATORY POLICIES AGAINST RADICALIZATION USE CASE

8.2 Intelligent Policies for The Denomination of Origin

There are several examples of open information that could be used for the Aragon use case, all the information comes from the Aragon open data initiative such as PAC, and SIGPAC open information. Link 1 in Table 3 provides information on the Geographical Information System for Agricultural Plots (SIGPAC), which makes it possible to identify geographically the plots declared by farmers and stockbreeders under any aid scheme relating to the area

cultivated or used by the animals. Viticultural registries information can be found at links 2, 3 and 4 of the same table.

Aragon Open Social Data is an application that captures and processes public information generated in the social networks of the institutional accounts. With this information a citizen control panel has been developed in which it is possible to know the existing conversation in and about Aragon offered and published by the institutional accounts. Aragon Open Social Data can be found at link 5 of Table 3.

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 3 – LINKS TO ARAGON USE CASE DATA STORES

8.2.1 Scenario 2.1 – Problem Statement

Analyze the impact in the form of comments on social networks, news on the wine brands of a designation of origin, after having carried out an advertising campaign on that designation of origin or at a general level on wine in Aragon.

8.2.1.1 Main Objective

Identify the impact of Aragon general marketing on different wine brands.

8.2.1.2 Key Performance Indicators

KPI 2.1.1	Search for the best impact vs money invested in marketing.

8.2.1.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Intelligent Policies for the Denomination of Origin	Scenario #2.1	Twitter, facebook, RSS feeds	Please refer to Table 3 earlier in the Use Case section.

TABLE 4 – DATA SOURCES LIST FOR SCENARIO 2.1 OF THE INTELLIGENT POLICIES FOR THE DENOMINATION OF ORIGIN USE CASE

8.2.2 Scenario 2.2 – Problem Statement

Analyze information extracted from social networks, web pages and blogs in order to discover new trends and generate new recommendations to a user of the PolicyCLOUD environment based on opinion analysis.

8.2.2.1 Main Objective

Identify the most interesting concepts, wines or elements about the wine world. Identify the influencers.

8.2.2.2 Key Performance Indicators

KPI 2.2.1	Identify trends in types of wines based on defined metrics.
KPI 2.2.2	Identify trends in agrifood based on defined metrics.
KPI 2.2.3	Identify influencers.

8.2.2.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Intelligent Policies for the Denomination of Origin	Scenario #2.2	Twitter, facebook, RSS feeds	Please refer to Table 3 earlier in the Use Case section.

TABLE 5 – DATA SOURCES LIST FOR SCENARIO 2.2 OF THE INTELLIGENT POLICIES FOR THE DENOMINATION OF ORIGIN USE CASE

8.2.3 Scenario 2.3 – Problem Statement

Perform Brand Analysis in the wine market, based on Information extracted from social networks, RSS channels and blogs.

8.2.3.1 Main Objective

Perform Brand Analysis on different brands of wines and in general of Denomination of Origin.

8.2.3.2 Key Performance Indicators

KPI 2.3.1	Quality of wines compared to price and taste.

8.2.3.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Intelligent Policies for the Denomination of Origin	Scenario #2.3	Twitter, facebook, RSS feeds	Please refer to Table 3 earlier in the Use Case section.

TABLE 6 – DATA SOURCES LIST FOR SCENARIO 2.3 OF THE INTELLIGENT POLICIES FOR THE DENOMINATION OF ORIGIN USE CASE

8.2.4 Scenario 2.4 – Problem Statement

We need to predict the quality of the next wine crop.

8.2.4.1 Main Objective

Predict the level of quality of the next wine crop by analyzing information from the Aragon region (climate, pests) and of the quality of the wine historically. Combine analyses made from data from twitter and other social media.

8.2.4.2 Key Performance Indicators

KPI 2.1.1	Quality of the next Wine Crop.

8.2.4.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Intelligent Policies for the Denomination of Origin	Scenario #2.4	Historic information from the Aragon region (climate, pests), and historic quality of wine	Please refer to Table 3 earlier in the Use Case section.

TABLE 7 – DATA SOURCES LIST FOR SCENARIO 2.4 OF THE INTELLIGENT POLICIES FOR THE DENOMINATION OF ORIGIN USE CASE

8.3 Urban Policy Making Through Analysis of Crowdsourced Data

8.3.1 Scenario 3.1 – Problem Statement

We need to visualize a Heatmap that maps the frequency of occurrence of incidents negatively affecting the road infrastructure and urban environment. Need to analyze the historical occurrence of incidents per location.

8.3.1.1 Main Objective

To validate existing city policies and reconstruction plans and to investigate if there is a need to adjust / update them or create new ones. To analyze the efficiency of past reconstruction/repairs.

8.3.1.2 Key Performance Indicators

KPI 3.1.1	Urban Environment / Road Infrastructure: Identify main problem areas. Identify tendencies (region, type, time to react). Increase impact of policies potential to proactively target areas for preventive measures.

8.3.1.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Urban Policy Making Through Analysis of Crowdsourced Data	Scenario #3.1	Sofia Municipality Contact Centre Signals from Citizens.	Sofia Municipality Contact Centre

TABLE 8 – DATA SOURCES LIST FOR SCENARIO 3.1 OF URBAN POLICY MAKING THROUGH ANALYSIS OF CROWDSOURCED DATA USE CASE

8.3.2 Scenario 3.2 – Problem Statement

We need to analyze environment and air quality data and incidents, incl. per area and time period (seasonality).

8.3.2.1 Main Objective

To validate existing city policies and plans and to investigate if there is a need to adjust/update them or create new ones. Assessment and adjustment of current initiatives, such as "Change your heater" and "Green ticket".

8.3.2.2 Key Performance Indicators

KPI 3.2.1	Air Quality: For the main problem areas identify patterns in terms of time, location, reason with defined metrics for pollution.
KPI 3.2.2	Correlation between pollution levels and municipal initiatives.

8.3.2.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Urban Policy Making Through Analysis of Crowdsourced Data	Scenario #3.2	Sofia Municipality Contact Centre Signals from Citizens.	Sofia Municipality Contact Centre
Urban Policy Making Through Analysis of Crowdsourced Data	Scenario #3.2	Airthings open data platform.	https://airthings.portal.azure-api.net/

TABLE 9 – DATA SOURCES LIST FOR SCENARIO 3.2 OF URBAN POLICY MAKING THROUGH ANALYSIS OF CROWDSOURCED DATA USE CASE

8.4 Predictive Analysis Towards Unemployment Risks Identification and Policy Making

8.4.1 Scenario 4.1 – Problem Statement

We need to utilize predictive analysis in order to anticipate common factors that relate to unemployment. Predictive analysis will use complex algorithms to predict future outcomes/trends which will assist policy makers in creating policies that can reduce unemployment.

8.4.1.1 Main Objective

The Identification of any trends or correlations that exist within unemployment data.

8.4.1.2 Key Performance Indicators

Factors such as age, gender and time-based statistics such as month/ year will be key indicators to highlight common trends in specific age groups. Summaries of the total amount of unemployed citizens within a specific time period can be a key component in identifying trends.

KPI 4.1.1	Total number of Male citizens unemployed
KPI 4.1.2	Total number of Female citizens employed
KPI 4.1.3	Total number of unemployed citizens divided in age ranges i.e. 16-24, 25-50, 50+

8.4.1.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Predictive Analysis Towards Unemployment Risks Identification and Policy Making	Scenario #4.1	The specified dataset includes statistics based on citizens claiming money to support themselves whilst they are unemployed.	https://opendata.camden.gov.uk/Business-Economy/Unemployment-Claimant-Count-LATEST/g3p6-usd3

TABLE 10 – DATA SOURCES LIST FOR SCENARIO 4.1 OF PREDICTIVE ANALYSIS TOWARDS UNEMPLOYMENT RISKS IDENTIFICATION AND POLICY MAKING USE CASE

8.4.2 Scenario 4.2 – Problem Statement

We need to conduct analysis on specific time periods. For example, unemployment is expected to go up during the year 2022 due to the current pandemic. Statistics recorded against the current year can help to identify the possible unemployment rate if there is a second wave of infections the following year.

8.4.2.1 Main Objective

Perform analysis based on set time periods such as month or year to help anticipate spikes therefore policy creators can take appropriate action.

8.4.2.2 Key Performance Indicators

KPI 4.2.1	Summary of unemployment figures based on a specific month
KPI 4.2.2	Summary of unemployment figures based on a specific year
KPI 4.2.3	Rate of % increase/decrease against previous year unemployment

8.4.2.3 Data Sources

Use Case	Scenario #	Data Source Description	Link(s)
Predictive Analysis Towards Unemployment Risks Identification and Policy Making	Scenario #4.2	The specified dataset includes statistics based on citizens claiming money to support themselves whilst they are unemployed.	https://opendata.camden.gov.uk/Business-Economy/Unemployment-Claimant-Count-LATEST/g3p6-usd3

TABLE 11 – DATA SOURCES LIST FOR SCENARIO 4.2 OF PREDICTIVE ANALYSIS TOWARDS UNEMPLOYMENT RISKS IDENTIFICATION AND POLICY MAKING USE CASE

9 Conclusion

The PolicyCLOUD Conceptual Model & Reference Architecture are presented in this report (Deliverable D2.2) released in M8 of the project. Updates of the deliverable will be published in M18 and M30.

The overall architecture was 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 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.

Several scenarios that have been prepared during specialized internal workshops discussing Architecture integration and end-to-end Use Case journey have been developed for the Use Cases of PolicyCLOUD and are included in this document. These scenarios will serve as a basis for end-to-end examples for the demonstration of the data ingest flow and data exploitation and for the analysis of the processing and data transformations along the complete data path. The detailed examples will be prepared during the completion of the first prototypes in M10 and will be included in the next update of Deliverable D2.2 in M18.

References

[1] International Organization for Standardization, "ISO/IEC/IEEE 29148:2011 – Systems and software engineering — Life cycle processes — Requirements engineering," ISO/IEC/IEEE, Nov. 2011.

[2] M. Stocker, M. Rönkkö and M. Kolehmainen, "Situational Knowledge Representation for Traffic Observed by a Pavement Vibration Sensor Network" in IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 4, pp. 1441-1450, Aug. 2014

[3] Stocker, M., Baranizadeh, E., Portin, H., Komppula, M., Rönkkö, M., Hamed, A., ... & Kolehmainen, M. (2014). Representing situational knowledge acquired from sensor data for atmospheric phenomena. Environmental Modelling & Software, 58, 27-47.

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

[5] Anderson, J. E. (1976). Cases in public policy-making. New York: Praeger.

[6] Setting Data Policies, Standards, and Processes, "http://www.mcpressonline.com/business-intelligence/setting-datapolicies-standards-and-processes.html"

[7] IDC, "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things", https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

[8] Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking, "http://www.data4policy.eu/state-of-the-art-report"

[9] Big Data: Basics and Dilemmas of Big Data Use in Policy-making, "http://www.policyhub.net/en/experience-andpractice/153"

[10] K. Moutselos, D. Kyriazis, V. Diamantopoulou and I. Maglogiannis, "Trustworthy data processing for health analytics tasks," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 3774-3779, doi: 10.1109/BigData.2018.8622449.

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

[12] Wirth, Rüdiger, and Jochen Hipp. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000.

[13] Jenkins open-source continuous integration, "https://jenkins-ci.org/"

[14] Puppet Labs: IT Automation Software, "https://puppetlabs.com/"

[15] Chef IT automation, "https://www.chef.io/chef/"

[16] A. Sheth, J. Larson. "Federated database systems for managing distributed, heterogeneous, and autonomous databases", ACM Comput. Surv. 22, 3 (September 1990), pp. 183-236, 1990.

[17] Data Gravity – in the Clouds, "http://blog.mccrory.me/2010/12/07/data-gravity-in-the-clouds/"

[18] What is Polyglot Persistence?, "http://www.jamesserra.com/archive/2015/07/what-is-polyglot-persistence/"

[19] Building a Just-In-Time Data Warehouse Platform with Databricks, "https://databricks.com/blog/2015/11/30/building-ajust-in-time-data-warehouse-platform-with-databricks.html"

[20] Apache Spark, "http://spark.apache.org/"

[21] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing", Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12), USENIX Association, Berkeley, CA, USA, 2-2, 2012.

[22] M. Armbrust, R.Xin, C. Lian, Y. Huai, D. Liu, J. Bradley, X. Meng, T. Kaftan, M. Franklin, A. Ghodsi, and M. Zaharia, "Spark SQL: Relational Data Processing in Spark", Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), ACM, New York, NY, USA, pp. 1383-1394, 2015.

[23] M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I Stoica, "Discretized streams: fault-tolerant streaming computation at scale",Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles (SOSP '13), ACM, New York, NY, USA, pp. 423-438, 2012.

[24] Structuring Spark: DataFrames, DataSets and Streaming, "https://spark-summit.org/east-2016/events/structuringspark-dataframes-datasets-and-streaming/"

[25] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things", Proceedings of the first edition of the MCC workshop on Mobile cloud computing, ACM, pp. 13-16, 2012.

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

[27] S. Chandrasekaran, O. Cooper, A. Deshpande, M. J. Franklin, J. M. Hellerstein, W. Hong, S. Krishnamurthy, S. R.Madden, F. Reiss, and M. A. Shah, "Telegraphcq: continuous dataflow processing", Proceedings of the 2003 ACM SIGMOD Int.Conference on Management of Data, pp. 668–668, ACM, 2003.

[28] D. Abadi, D. Carney, U. Çetintemel, "Aurora: a new model and architecture for data stream management", The VLDB Journal - The Int. Journal on Very Large Data Bases, 12(2):120–139, 2003.

[29] D. Abadi, Y. Ahmad, M. Balazinska, "The design of the borealis stream processing engine", In CIDR, volume 5, pp. 277– 289, 2005.

[30] M. Cherniack, H. Balakrishnan, M. Balazinska, D. Carney, U. C¸ etintemel, Y. Xing, and S. B. Zdonik, "Scalable distributed stream processing", CIDR, 2003.

[31] A. Shah, J. M. Hellerstein, S. Chandrasekaran, and M. J. Franklin, "Flux: An adaptive partitioning operator for continuous query systems,", ICDE, pp. 25–36, 2003.

[32] V. Gulisano, R. Jimenez-Peris, M. Patiño-Martinez, C. Soriente, P. Valduriez, "StreamCloud: An Elastic and Scalable Data Streaming System", IEEE Transactions on Parallel and Distributed Processing (TPDS), Vol 23, issue 12. pp: 2351-2365, December 2012.

[33] Apache Storm, "http://storm.apache.org/"

[34] Esper, "http://www.espertech.com/products/esper.php"

[35] INDIGO PaaS Orchestrator, "https://indigo-paas.cloud.ba.infn.it/home/login"

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

[37] Moleculer – Progressive Microservices framework for Node.js, https://moleculer.services/

[38] Data Specifications, "http://inspire.ec.europa.eu/data-specifications/2892"

[39] HL7, "www.hl7.org/"

[40] Interfaceware HL7 Overview, "www.interfaceware.com/hl7.html"

[41] Cordis, "http://cordis.europa.eu/cerif"

[42] Incubator, "https://www.w3.org/2005/Incubator/ssn/ssnx/ssn"

[43] Open Geospatial Standards, "www.opengeospatial.org/standards/sensorml"

[44] FOAF, "http://www.foaf-project.org/"

[45] SIOC project, "http://sioc-project.org/"

[46] Tom Heath and Christian Bizer, "Linked Data: Evolving the Web into a Global Data Space (1st edition)". Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

[47] JSON, "http://json-ld.org/"

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

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

[50] The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages, IEEE Computer Society.

[51] Maceachren, A. M. How Maps Work: Representation, Visualization, and Design, The Guilford Press, 2004.

[52] Aigner, W., S. Miksch, et al. Visualizing time-oriented data - A systematic view, Computer Graphics 31(3): 401-409, 2007.

[53] Kwan, M.-P., Space-Time Paths, Madden, M. (ed.) Manual of Geographic Information Systems. American Society for Photogrammetry and Remote Sensing, 2009.

[54] Wood, Jo, Aidan Slingsby, and Jason Dykes. "Using treemaps for variable selection in Spatio-Temporal visualization." Information Visualization 7.3: 4, 2008.

[55] Andrienko, N. & Andrienko, G., Spatial Generalization and Aggregation of Massive Movement Data, IEEE Transactions on Visualization and Computer Graphics, Vol. 17, No. 2, 2011.

[56] Boyandin, I.; Bertini, E.; Bak, P. & Lalanne, D.: Flowstrates: An Approach for Visual Exploration of Temporal OriginDestination Data. Computer Graphics Forum, Vol. 30, No. 3, 2011.

[57] Eccles, R.; Kapler, T.; Harper, R. & Wright, W., Stories in GeoTime, Information Visualization, Vol. 7, No. 1, 2008.

[58] Fuchs, G. & Schumann, H., Visualizing Abstract Data on Maps, Proceedings of the International Conference Information Visualisation (IV), IEEE Computer Society, 2004.

[59] Shanbhag, P.; Rheingans, P. & desJardins, M. Temporal Visualization of Planning Polygons for Effcient Partitioning of Geo-Spatial Data. Proceedings of the IEEE Symposium on Information Visualization (InfoVis), IEEE Computer Society, 2005.

[60] Fabrikant, S. and D. Josselin La "cartactive", cartographie en mouvement:un nouveau domaine de recherche pluridisciplinaire ou un plan de la géomatique? Revue Internationale de Géomatique, Numéro spécial "cartographie animée et interactive" 13(1): 7-14. 2003.

[61] Tominski, C., P. Schulze-Wollgast, et al. 3D information visualization for time dependent data on maps, 2005.

[62] Harrower, M. and S. Fabrikant, Eds. (2008). The role of map animation in geographic visualization. Geographic Visualization. Chichester UK, Wiley and Sons, 2008.

[63] Johnson, I. Indexing and Delivery of Historical Maps online using TIMEMAP. National Library of Australia Magazine, 2005.

[64] Pan, J., & McElhannon, J. Future edge cloud and edge computing for internet of things applications. IEEE Internet of Things Journal, 5(1), 439-449, 2018.

[65] P. Pietzuch, J. Ledlie, J. Shneidman, M. Roussopoulos, M. Welsh, and M. Seltzer, "Network-Aware Operator Placement for Stream-Processing Systems", 22nd International Conference on Data Engineering, ICDE '06, pp. 49–. IEEE Computer Society, 2006.

[66] V. Cardellini, V. Grassi, F. Lo Presti, and M. Nardelli, "Distributed QoS-aware Scheduling in Storm", 9th ACM International Conference on Distributed Event-Based Systems, pp. 344-347, ACM, 2015.

[67] Y. Xing, S. Zdonik, and J.-H. Hwang, "Dynamic Load Distribution in the Borealis Stream Processor", 21st International Conference on Data Engineering, ICDE '05, pp. 791–802, IEEE Computer Society, 2005.

[68] M. Hirzel, R. Soule, S. Schneider, B. Gedik, and R. Grimm, "A Catalog of Stream Processing Optimizations", ACM Computing Surveys, 46(4):1–34, Mar. 2014.

[69] Mijumbi, Rashid, et al. "Management and orchestration challenges in network functions virtualization." IEEE Communications Magazine 54.1: 98-105, 2016.

[70] SIIA White Paper, "Data-Driven Innovation – A Guide for Policymakers: Understanding and Enabling the Economic and Social Value of Data", 2013.

[71] O. Teme, J. Polonetsky, "Big Data for All: Privacy and User Control in the Age of Analytics", Northwestern Journal of Technology and Intellectual Property, 2012.