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

D2.1 STATE OF THE ART & REQUIREMENTS ANALYSIS

Dissemination Level	PU
Due Date of Deliverable	30/06/2020 (M06)
Actual Submission Date	30/06/2020 (M06)
Work Package	WP2, Requirements, Architecture & Innovation
Task	T2.1 Requirements Elicitation & State of the Art Analysis
Type	Report
Approval Status	
Version	V1.3
Number of Pages	p.1 – p.127

Abstract: This document contains an initial analysis of the state of the art of the baseline technologies that will be used in the scope of the project. Moreover, the elicitation of the requirements coming both from the use cases and the technological partners is included, along with the list of business goals and requirements.

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

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
0.0	17/02/2020	Initial ToC	LXS
0.1	17/02/2020	Add preliminary input on Section 1, 3, 4	LXS
0.2	18/02/2020	Add preliminary input on Section 5, 6, 7, 8, 9	LXS
0.3	25/03/2020	Input in use cases, dataset definition and SotA	ALL
0.4	31/03/2020	Input in use cases, dataset definition and SotA	ALL
0.5	13/04/2020	Merge all inputs, and input on executive summary, conclusions and overall QA	LXS
0.6	13/04/2020	Submitted for internal release	LXS
0.7	13/04/2020	New integrated version for M06	LXS
0.8	20/04/2020	Input for LON use case	LON
0.9	02/06/2020	Input for requirements	ALL
1.0	23/06/2020	Document ready for internal review	LXS
1.1	24/06/2020	1st review	UPRC
1.2	25/06/2020	2nd review	UBI
1.3	30/06/2020	Quality check. Document ready for submission	UPRC

Author List

Organisation	Name
LXS	Sandra Ebro, Boyan Kolev, Jose María Zaragoza, Patricio Martinez, Luis Miguel Garcia
IBM	Ofer Biran
ICCS	Konstantinos Moutselos
ATOS	María Ángeles Sanguino, Jorge Montero, Ana Luiza Pontual, Miquel Milà, Ricard Munné
EGI	Giuseppe La Rocca
UPRC	George Manias, Argyro Mavrogiorgou, Athanasios Kiourtis, Ilias Maglogiannis, Nikitas Sgouros
ITA	Rafael del Hoyo
UBI	Giannis Ledakis, Konstantinos Theodosiou
MAG	Armend Duzha, Nikos Achilleopoulos
SOF	Petya Nikolova, Iskra Yovkova
ITA	Vega Rodrigálvarez
SARGA	Javier Sancho

Abbreviations and Acronyms

Abbreviation/Acronym	Definition
EBPM	Evidence Based Policy Making
EC	European Commission
EOSC	European Open Science Cloud
KPI	Key Performance Indicator

Contents

List of Tables

List of Figures

Executive Summary

This is the first of the series of deliverables that specify the use case scenarios, their involved datasets and their relevant user requirements as well as the system and technical requirements that are being imposed by the platform. The purpose of these series is to track those requirements throughout the project and update them during the progress of the project. The approach that is followed is twofold: A top-down approach that is followed with respect to the user requirements that were collected by the use case providers themselves, after specifying the business goals and objectives of the use case, along with a concrete definition of the scenario. Moreover, a bottom-up approach is additionally complemented that aims to identify and analyse the technical requirements with respect to the technical work packages that are focusing on the platform technological needs.

The result of this analysis is a list of measurable unambiguous requirements that will drive the design of the overall architecture of the PolicyCLOUD platform, focusing on serving all different needs of the various use cases of the project. As the project will progress, updated versions of this deliverable will help the architecture designers of the platform and its software developers to adjust the overall architecture and its implementation accordingly, with respect to the principles of the agile methodology. Moreover, in order for the platform to keep track with the latest technological advances, a state-of-the-art analysis has been performed regarding the major technologies that are envisioned to be exploited, along with a list of several projects whose technological assets might be candidate to be incorporated in the overall solution.

This deliverable has been released on M06 of the project, and its main aim is to specify the basic scenarios of the use cases, which drove the user requirements based on their perspective, and the initial technical requirements as foreseen by the technical partners that are being involved in the design of the overall architecture of the platform. Updated versions of this document will be released on M12 and M22 respectively.

1 Introduction

This document purpose is to provide an initial list of measurable and specific **user, business and system requirements** that will drive the design of the architecture of the PolicyCLOUD platform and will be used as the basis for the implementation of the relevant functionalities that will be offered by the various software components of the platform.

This report is the first deliverable of those that need to be produced in the context of the work that will be carried out in PolicyCLOUD's task T2.1 "*Requirements Elicitation & State of the Art Analysis*", whose main goal of this task is to collect the user and system requirements and tracked during the course of the project. The analysis of the requirements will produce **a measurable and unambiguous requirement set**, which will be tracked against the developments during the project lifecycle in order to ensure that the PolicyCLOUD complexity will be fully addressed and properly considered. Moreover, another important goal of task T2.1 is **to investigate and analyse the State-of-the-Art (SotA)** for PolicyCLOUD technologies. Both these two goals will be a valuable input for the design of the overall platform architecture and all research activities of the project. As the importance of this task is high, as it affects the overall design of the project, the task started in M01, and an internal report that was used of internal purposes was already produced on M03. At that moment, some initial information was already collected in order for the technical members of the consortium to be able to kick off their research tasks. The initial version of that report was further refined during the next period and this is the first official version of the document that reflects that has been done until M06. As this task duration ends in M22, there will be two additional versions of this document that will refine the deliverable and are expected to be produced in the upcoming months. Their names will be:

- D2.1.2: State of the art and requirements analysis II (M12)
- D2.1.3: State of the art and requirements analysis III (M22)

The analysis and elicitation of the requirements have been carried out taking into account the exact needs and concerns that have been identified by the current communities, end-users and related actors that are related to the PolicyCLOUD use cases and providers of the corresponding technologies. As a result, the analysis that has been made not only specifies **use case requirements**, that can be also considered as *stakeholder requirements* by ISO/IEC/IEEE 29148:2011 [1], but also **technical requirements** that can be considered as *system* and *software requirements*. At the time that this version of the deliverable is published (M06) it is very early for all the system requirements to be identified, and therefore, the main focus was given on the definition of the use case requirements, the analytical description of each case, the types of different datasets that will be brought to the platform and are considered to be used, while on the same time, main focus was additionally given on the definition of the software technology requirements that are being imposed by each of the main components. Due to the complexity of the overall architecture and the wide variety of the technologies that will formulate the envisioned PolicyCLOUD platform, system requirements that sit in the middle of the user and software ones, will be further developed in the second version of this deliverable.

Apart from the list of the functional and non-functional requirements, this deliverable was planned to have an additional section where it would have also described the **various categories of the stakeholder of the PolicyCLOUD platform, the different envisioned business model and the expected business outcomes and business goals**. This analysis, even if it is not addressing strictly technical perspectives of the project, it was considered to be a valuable input for the corresponding tasks of WP7, mainly on what concerns the project's road mapping & business development. However, it the work that has been carried in this task in correspondence with the WP7 was decided to be included in the D7.2 deliverable. Moreover, in order to better understand the software technology requirements, this deliverable includes in section 1 an analysis of the state-of-the-art related technologies. At the same time, an initial and non-exhaustive list of relevant research initiatives and projects has also been provided, along with the description of the baseline technologies that the technical partners will bring to the project.

This document is organized as follows: Section 1.1 explains the requirements engineering method that has been followed; Section 1 provides an analytical description of the use cases, along with the initial list of the user requirements, while section 0 describes the various datasets that each use case is intending to use, along with all possible data regulatory constraints. Section 1 defines the different roles of the users of the PolicyCLOUD ecosystem, while sections 1 and 0 provide the technical requirements of the platform. Sections 1 and 0 provide the state-of-the-art analysis and specify a list of the baseline technologies that are intended to be used in the development and implementation of the platform, while section 0 finally concludes the document.

1.1 Method

The engineering method to gather all user and technical requirements for the PolicyCLOUD project follows the ISO/IEC/IEEE 29148:2011 norm, as already mentioned in the previous subsection. This norm describes two main processes and practices that need to be executed in an iterative and recursive manner.

The first process is related to the **definition of the requirements coming from the stakeholders**. Its purpose is to define the requirements for a system that can provide the services needed by users and other stakeholders in a defined environment. The output of this process is the *Stakeholder Requirements Specification (StRS)*. On the other hand, the second process that is defined in the norm is related with the **requirements analysis** whose main purpose is to transform the stakeholder, requirement-driven view of desired services into a technical view of a required product that could deliver those services. The outcome of the second process is *the System Requirements Specification (SyRS)* and the *Software Requirements Specification (SRS).*

The specification of the overall requirement can be provided in three levels of detail, which serve as input to different practices or stages in the architectural design process. Their relations can be seen at Figure **1**. Moreover, with respect to the ISO/IEC/IEEE 29148:2011, Table **1** describes the relations of each of the three outcomes of the two processes that are defined: the Stakeholder Requirements Specification (StRS), the System Requirements Specification (SyRS) and the Software Requirements Specification (SRS), including the architecture domain whose decisions are informed by them.

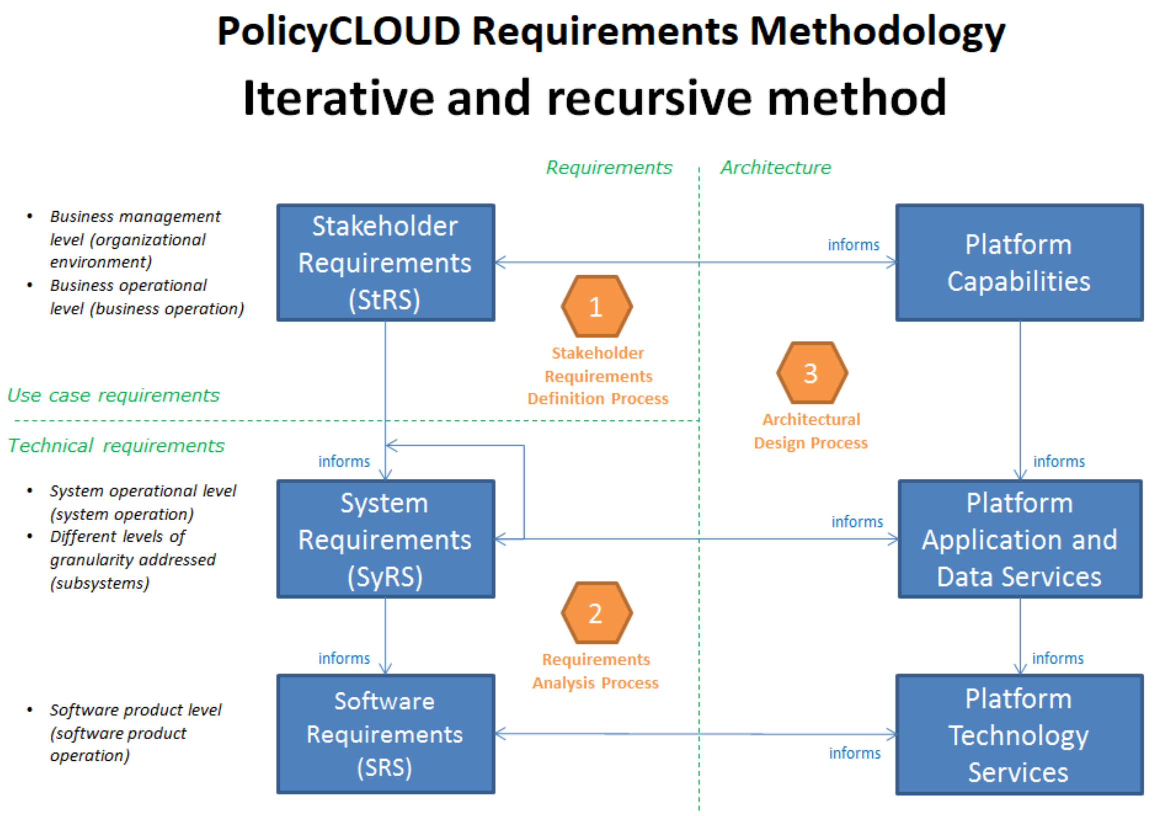

PolicyCLOUD Requirements Methodology
Iterative and recursive method
Requirements
Architecture
• Business management level (organizational environment)
• Business operational level (business operation)
Stakeholder Requirements (StRS)
informs
Platform Capabilities
1
Stakeholder Requirements Definition Process
3
Architectural Design Process
Use case requirements
Technical requirements
informs
informs
• System operational level (system operation)
• Different levels of granularity addressed (subsystems)
System Requirements (SyRS)
informs
Platform Application and Data Services
2
Requirements Analysis Process
informs
informs
• Software product level (software product operation)
Software Requirements (SRS)
informs
Platform Technology Services

FIGURE 1: METHODOLOGY FOR REQUIREMENTS ENGINEERING

Work Product	Acronum	Description	Informed Architecture Domain
Business Requirements Specification	StRS	This contains the requirements as defined by the use case providers	Platform Capabilities (business architecture)
System Requirements Specification	SyRS	This defines the platform level requirements	Platform Applications and Data Services Architecture
Software Requirements Specification	SRS	This contains the specific requirements of each one of the individual components	Platform Technology Architecture

TABLE 1: WORK PRODUCTS DESCRIPTION

Moreover, in order to highlight the key business requirements so as to indicate the implied technical requirements for the overall architecture of the PolicyCLOUD platform, we used the *TOGAF® Series Guide : Business Scenarios*[1] methodology which facilitates the identification of the requirements from the stakeholders' point of view. This technique aims to validate, elaborate and modify the premise behind an architecture effort, by

[1] https://publications.opengroup.org/g176

focusing on the understanding and documentation of the key elements of a business scenario using iterations in an agile manner.

Finally, to better formalize the requirements, we use the following attributes:

- **Level of detail**: Following the use of ISO/IEC/IEEE 29148:2011, we use the following levels: Stakeholder, System and Software (i.e., technology details).
- **Type**: Types of requirements are functional: FUNC (function), DATA (data); and non-functional: L&F (Look and Feel Requirements), USE (Usability Requirements), PERF (Performance Requirements), ENV (Operational/Environment Requirements), and SUP (Maintainability and Support Requirements).
- **Priority**: Requirements can have different priorities: MAN (mandatory requirement), DES (desirable requirement), OPT (optional requirement), ENH (possible future enhancement).

2 Use case Requirements and Scenarios

The purpose of this section is to present the business usage scenarios along with the list of the initial requirements that have been defined by each of the four use cases of the PolicyCLOUD project. These requirements formulate the list of the overall Stakeholder Requirements, according to the engineering methodology that has been analysed in Table **1**.

Each of the four cases describes the exact usage from a use case perspective at a high-level description. It is worth to be mentioned that the complete definition of the detailed scenario that is needed, was not the focus of this analysis, as this is part of the work to be carried out in the scope of T6.2 "*Use case Definition & Design".* The scope of the work that is being reported in this section is rather the general descriptions that are more related with the general definition of the behaviour and identification of the important necessities that the architecture should comply with, so that they can be taken into account from the very beginning of the project. In any case, there has not been the necessity for the complete definition of the use case in this section and at this version of the document, as the project has two additional iterations to upgrade the requirements and refine them accordingly, in correspondence with the work that will have to be carried out in T6.1. The list of the Stakeholder Requirements and the general description of the use case, however, can provide an overview on the main behavioural patterns involving the different actors and aims to define and align the initial design of the architecture (D2.2). The descriptions of the scenarios are complemented with UML Use Case Diagrams in order to identify the different actors, prerequisites and the description of the behaviour.

The following subsections firstly give an introductory overview of the purpose of each scenario, followed by a more detailed description of the use case. Then, the description of the different user stories that formulate each scenario is presented, along with the corresponding UML diagram, and finally the initial list of the stakeholder requirements is reported.

2.1 UC#1: Participatory policies against radicalization

2.1.1 Goals and Objectives

UC#1 aims to develop a collaborative data-driven application for the validation of existing policies to counter radicalization based on a participatory review of data coming from social media and open datasets. In addition, it will provide useful hints to policy makers at local, regional and national level to adjust / update the current policies and investigate whether new ones are needed.

This use case will address the challenge of radicalisation by offering policy makers ICT-based tools for enabling them to **monitor, identify, analyse, visualize and predict potential risks of radicalization**, while at the same time **allow them to interact with other stakeholders** (i.e. data analytics professionals, social scientists, legal experts) during the creation and modelling of policies and specific measures against counter-violent extremism.

UC#1 will adopt the PolicyCLOUD technologies developed during the project, and in particular:

- The **PolicyCLOUD opinion mining and sentiment analysis** tools, which provides
    - social media analysis towards the identification of radicalization activities and actors (individual, groups) involved and linking of data about terroristic groups and attacks with radicalization efforts.
        - Alleviate the negative consequences of counter-radicalisation policies (e.g. restrictions) by making them more targeted.

- big data analytics to identify origins of radicalization efforts (including countries/regions and terroristic groups conducting them), risk assignment probabilities to suspects of radicalization efforts and segmentation of radicalization efforts and subjects on the basis of demographics and risks.
- opinion / sentiment classification and user type classification, as well as pattern identification and analysis over time will also be enabled through approaches developed in this task.

- The **PolicyCLOUD situational knowledge acquisition and analysis** tools, which provides:
    - knowledge acquisition from real-world data using statistical algorithms and machine learning techniques in combination with collective knowledge (out of the clusters / collections of policies – Task 5.5), and predictive risk analysis. The knowledge derived will influence different types of proposed decisions towards target communities given the developed knowledge regarding efficiency of decisions, current status and policy planning.
- The **PolicyCLOUD visualization** technologies which will:
    - enable the policy maker to depict visually the radicalisation trends and poles.

2.1.2 Description of Scenarios

The following tables define the various scenarios of this use case.

Section	Description
ID	SCE-PPR-01
Title	Social Media Configuration
Description	Insert new keywords to search
Actors	Administrator, Policy makers
Objectives	Describing elements from social networks where information has to be analysed
Pre-Conditions	● The Administrator should have been identified on the platform. ● Social networks used: Twitter, Facebook, Reddit, RSS Feed
Process Description	● Administrator selects "My research interests" ● Systems shows a wizard where the user enters configuration parameters:social networks, topics to inspect. ● Administrator inserts, modifies or deletes social networks and keywords ● System saves configuration and add new parameters to be used by the probe
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 2: SCENARIO SCE-PPR-01

Section	Description
ID	SCE-PPR-02
Title	Social Media Analysis
Description	Identify radicalization efforts and the actors (individual, groups) involved
Actors	Data Collector, Analyst, Policy maker
Objectives	• Identify Social network's users and groups on several applications • Identify new keywords, semantics on Hashtags • Identify comments on posts
Pre-Conditions	Verify activity on platforms: Twitter, Reddit, RSS Feed
Process Description	N/A
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 3: SCENARIO SCE-PPR-02

Section	Description
ID	SCE-PPR-03
Title	Assessment of social media observations against perceived radicalisation efforts
Description	Linking of data about terroristic groups and attacks with radicalization efforts
Actors	Sector Analyst, Policy makers
Objectives	• Observer social network's posts • Identify comments on posts • Observe social interactions
Pre-Conditions	Observe activity on platforms, Observe events and social attitude about radicalization efforts
Process Description	Data and Semantics linking
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 4: SCENARIO SCE-PPR-03

Section	Description
ID	SCE-PPR-04
Title	Open Datasets Configuration
Description	Insert new keywords to search

Section	Description
Actors	Administrator, Policy makers
Objectives	Describing elements from open datasets where information has to be analysed
Pre-Conditions	• The Administrator should have been identified on the platform. • Datasets used: GTD[2]
Process Description	• Administrator selects "My research interests" • Systems shows a wizard where the user enters configuration parameters: topics to inspect • Administrator inserts, modifies or deletes keywords • System saves configuration and add new parameters to be used by the probe
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 5: SCENARIO SCE-PPR-04

Section	Description
ID	SCE-PPR-05
Title	Open Datasets Analysis
Description	Identify radicalization efforts and the actors (individual, groups) involved
Actors	Data Collector, Analyst, Policy makers
Objectives	• Identify potential users and groups based on several observations • Identify new keywords
Pre-Conditions	Verify activity on the open datasets: GTD
Process Description	N/A
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 6: SCENARIO SCE-PPR-05

Section	Description
ID	SCE-PPR-06
Title	Visual Representation of Radicalization Trends and Poles
Description	Depict visually the radicalisation trends and poles through a heatmap
Actors	Sector Analyst, Policy maker
Objectives	• Identify users and groups based on their location • Aggregate users and groups based on attack types

[2] https://www.start.umd.edu/gtd

Section	Description
Pre-Conditions	Access to data: GTD
Process Description	Patterns of behaviour are often documented in a narrative form, this scenario give useful hints which will help policy makers in identifying patters base of demographic, location, attack types etc.
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 7: SCENARIO SCE-PPR-06

Section	Description
ID	SCE-PPR-07
Title	Visual Representation of Radicalization Trends and Poles
Description	Make policies against radicalization and violent extremism more transparent and open to public scrutiny through the use of the proposed technologies that create and structure open data datasets with statistics about radicalization efforts in social media and provide APIs and visualization tools for accessing them i.e. making them easily accessible and reusable by third parties
Actors	Sector Analyst, Policy makers
Objectives	N/A
Pre-Conditions	• Ensure continuous interactions and collaboration with relevant stakeholders at any level (local, regional, national and EU)
Process Description	N/A
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the second release of this deliverable

TABLE 8: SCENARIO SCE-PPR-07

2.1.3 Stakeholder Requirements

The following tables contain the initial list of the stakeholder requirements for the scenarios of this use case that were described in the previous subsection.

Section	Description
ID	REQ- PPR-01
Title	Restricted access
Level of detail	User
Type	FUNC
Description	A username and password are required to configure which information have to be gathered and analyses and hierarchical categorization
Additional Information	N/A

Section	Description
Actor	Administrators, users
Priority	MAN
Reference Use Case	SCE-PPR-01, SCE-PPR-04
Success Criteria	Nobody without login/password can access to application
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 9: STAKEHOLDER REQUIREMENT REQ-PPR-01

Section	Description
ID	REQ-PPR-02
Title	Opinion Mining
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks should be analysed in order to better understand what individuals or groups are saying about a specific discussion topic
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-02, SCE-PPR-03, SCE-PPR-05
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 10: STAKEHOLDER REQUIREMENT REQ-PPR-02

Section	Description
ID	REQ-PPR-03
Title	Sentiment Analysis
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks and open datasets should be analysed in order to know how individuals or groups feel about a specific discussion topic and capture their feelings
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-02, SCE-PPR-03, SCE-PPR-05
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document

Section	Description
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 11: STAKEHOLDER REQUIREMENT REQ-PPR-03

Section	Description
ID	REQ-PPR-04
Title	Text classification
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks and open datasets should be grouped in predefined "clusters"
Additional Information	Classification must be agreed with policy regulators
Actor	Administrator, Policy Makers
Priority	MAN
Reference Use Case	SCE-PPR-02, SCE-PPR-03, SCE-PPR-05
Success Criteria	All texts have to be tagged into one or more clusters
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 12: STAKEHOLDER REQUIREMENT REQ-PPR-04

Section	Description
ID	REQ-PPR-05
Title	Extraction of entities
Level of detail	Stakeholder
Type	FUNC
Description	More relevant entries (age range, location (city area), attack types etc.) should be extracted from gathered data
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-03, SCE-PPR-05, SCE-PPR-07
Success Criteria	Entities extracted from texts
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 13: STAKEHOLDER REQUIREMENT REQ- PPR-05

Section	Description
ID	REQ-PPR-06
Title	Personal Data

Section	Description
Level of detail	Stakeholder
Type	DATA
Description	Personal data like names, address, will not be stored
Additional Information	Full compliance with GPDR and national laws
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-03
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 14: STAKEHOLDER REQUIREMENT REQ- PPR-06

Section	Description
ID	REQ-PPR-07
Title	Twitter information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from tweets: text, user (alias), channel, location, date, origin source
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-03
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 15: STAKEHOLDER REQUIREMENT REQ-PPR-07

Section	Description
ID	REQ-PPR-08
Title	Facebook information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from posts and comments in group discussions: text, user (alias), location, attack type, etc
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	MAN

Section	Description
Reference Use Case	SCE-PPR-03
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 16: STAKEHOLDER REQUIREMENT REQ-PPR-08

Section	Description
ID	REQ-PPR-09
Title	Reddit Information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from Reddit posts: text, user (alias), location, date, origin source
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	DES
Reference Use Case	SCE-PPR-03
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 17: STAKEHOLDER REQUIREMENT REQ-PPR-09

Section	Description
ID	REQ-PPR-10
Title	RSS & web pages
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from web pages: text, source, date, title
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-03
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 18: STAKEHOLDER REQUIREMENT REQ-PPR-10

Section	Description
ID	REQ-PPR-11
Title	Information gathering from open datasets
Level of detail	Stakeholder
Type	DATA
Description	Data from different open sources will be gathered
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-05
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 19: STAKEHOLDER REQUIREMENT REQ-PPR-11

Section	Description
ID	REQ-PPR-12
Title	Data Analysis in near real-time
Level of detail	Stakeholder
Type	PERF
Description	Information collected from different dataset should be analysed every predefined time
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-PPR-03, SCE-PPR-05
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 20: STAKEHOLDER REQUIREMENT REQ-PPR-12

Section	Description
ID	REQ-PPR-13
Title	Visualization
Level of detail	Stakeholder
Type	L&F
Description	Dashboard should show more relevant information at a glance
Additional Information	N/A
Actor	End User, Policy Maker

Section	Description
Priority	DES
Reference Use Case	SCE-PPR-06, SCE-PPR-07
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 21: STAKEHOLDER REQUIREMENT REQ-PPR-13

Section	Description
ID	REQ-PPR-14
Title	Risk Prediction
Level of detail	Stakeholder
Type	L&F
Description	Dashboard should show predictions on potential risks / threats and their location
Additional Information	N/A
Actor	End User, Policy Maker
Priority	DES
Reference Use Case	SCE-PPR-06, SCE-PPR-07
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 22: STAKEHOLDER REQUIREMENT REQ-PPR-14

Section	Description
ID	REQ-PPR-15
Title	Working hours
Level of detail	Stakeholder
Type	SUP
Description	Gathering & analysis information should be working 24x7 Web page should be available 24x7
Additional Information	
Actor	
Priority	DES
Reference Use Case	SCE-PPR-01
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 23: STAKEHOLDER REQUIREMENT REQ-PPR-15

2.2 UC#2: Intelligent policies for the development of denomination of origin

2.2.1 Goals and Objectives

The main objective of this use case is to address **policies driving Denomination of Origin** through the use of **PolicyCLOUD analytics technologies**.

The most important ideas of the use case are:

- Identification of heterogeneous data sources like social networks (Twitter, Facebook, Instagram), open data sources, any other documents provided by the Government, industries and etc., to extract hidden patterns and information.
- Extracting, analysing and classifying information based on defined ontologies to generate reports about the state of the art, news about the different products covered by the denomination of origins in Aragón
- Knowing deeply which are the new trends on international markets, emerging issues about the product covered, policy recommendation
- Brand analysis of Aragón's OD and the competence, and recommendations to implement policies
- Creating KPIs to analyse the status of denomination of Origin.
- Creating and implementing new policies to help to create proposals of differential value on the agri-food sector in Aragon which helps to the specialization and development of Denomination of Origin.
- Evaluating the impact of implemented policies and comparison with the older ones.

Taking into account these objectives, the purpose of the use case is to create a tool that allows Governments analyse what is happening and design and improve policies around the Denomination of Origin in Aragon.

2.2.2 Description of Scenarios

The following tables define the various scenarios of this use case.

Section	Description
ID	SCE-IIPDD-01
Title	Configuration Panel
Description	Identifying and configuring data sources to analyse: • Social network's users and communities • Key words to search on social networks • Information provided from news webs and blogs Configuration of categories to classify information provided from different channels
Actors	Administrator, Policy makers
Objectives	Describing elements in order to identify relevant information
Pre-Conditions	1. The Administrator should be identified on the platform. 2. Social networks used: Twitter, Facebook, LinkedIn and Instagram 3. Other channels: news pages, Blogs
Process Description	1. <<Include User Identification>> Administrator enters login and password 2. System presents different options to select under the control panel 3. Administrator selects the option to configure 4. System shows configuration options 5. Administrator configure data sources 6. Systems save configuration

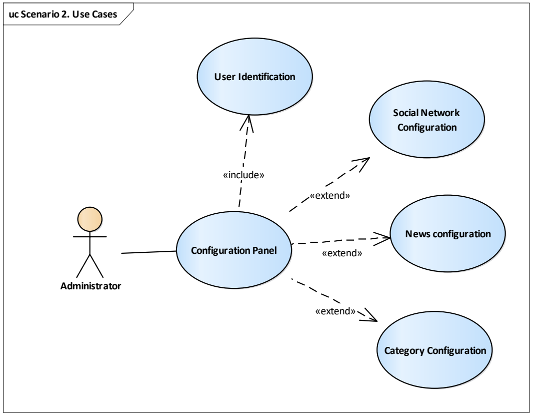
Section	Description
Variations	1a. If user's login or password is wrong, they will not be able to access to the control panel 3a. <<Extends Social Network configuration>> 3b. <<Extends News Configuration>> 3c. <<Extends Category Configuration>>
Post-Conditions	1. New sources are identified and analysed 2. Hierarchical text categorization defined (taxonomy or ontology)
UML User Case Diagram	 uc Scenario 2. Use Cases User Identification Social Network Configuration «include» «extend» News configuration Configuration Panel «extend» Administrator «extend» Category Configuration

TABLE 24: SCENARIO SCE-IIPDD-01

Section	Description
ID	SCE-IIPDD-02
Title	Social Network Configuration
Description	Inserting new users, communities & words to search
Actors	Administrator, Policy makers
Objectives	Describing elements from social networks where information has to be analysed
Pre-Conditions	1. The Administrator should be identified on the platform. 2. Social networks used: Twitter, Facebook, LinkedIn and Instagram
Process Description	1. Administrator selects "Social Networks Configuration" 2. Systems shows a wizard where administrator enters configuration parameters: social networks, user or community to follow, topics to inspect from Aragon's DO… 3. Administrator inserts, modifies or deletes social network configuration 4. System saves configuration and add new parameters to be used by the probe
Variations	
Post-Conditions	
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 25: SCENARIO SCE-IIPDD-02

Section	Description
ID	SCE-IIPDD-03
Title	News configuration
Description	Adding new news pages, channels or blogs to extract relevant information
Actors	Administrator, Policy makers
Objectives	Configuration of the system in order to extract relevant information about denominations of origin and related subjects to create and improve policies

Section	Description
Pre-Conditions	1. The Administrator should be identified on the platform. 2. Text channels: news pages, Blogs
Process Description	1. Administrator selects "News configuration" 2. Systems shows a configuration panel where the user defined the source of the document and the type. 3. System saves configuration and add new parameters to be used by the probe
Variations	N/A
Post-Conditions	N/A
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 26: SCENARIO SCE-IIPDD-03

Section	Description
ID	SCE-IIPDD-04
Title	Category Configuration
Description	Definition of a set of hierarchical categories which allow the system to classify information taken from different sources.
Actors	Administrator, Policy makers
Objectives	Defining a category for the use case
Pre-Conditions	1. The Administrator should be identified on the platform.
Process Description	1. Administrator selects "Category Configuration" 2. Systems shows a wizard where administrator define a category and the elements around it 3. Administrator inserts, modifies or deletes information related to categories 4. System saves configuration and add new parameters to be used by analyzers
Variations	
Post-Conditions	1. New categories and words are included in the system
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 27: SCENARIO SCE-IIPDD-04

Section	Description
ID	SCE-IIPDD-05
Title	Dashboard
Description	Displaying information which helps to policy maker to take the right decision related to Denomination of Origin subjects
Actors	Policy Maker, End user
Objectives	Providing support to policy makers
Pre-Conditions	
Process Description	1. System presents different options to select under the dashboard 2. User selects the option to configure 3. System shows a panel with obtained results and recommendations to the user
Variations	2a. Categorization 2b. Trending 2c. Brand analysis

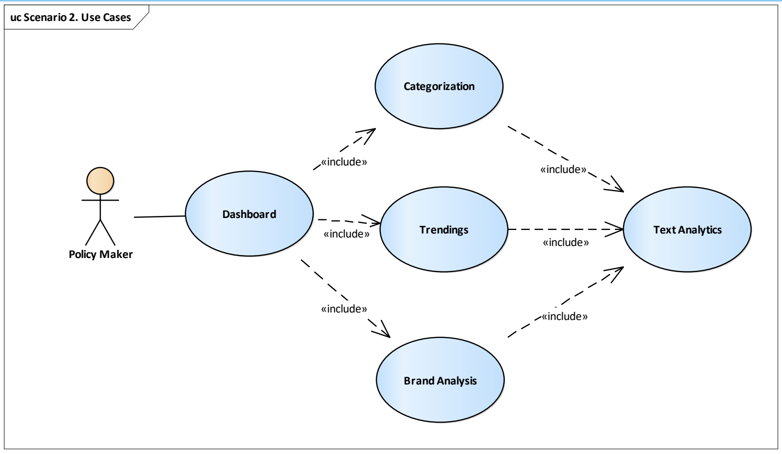
Section	Description
Post-Conditions	N/A
UML User Case Diagram	 uc Scenario 2. Use Cases Categorization «include» «include» Policy Maker Dashboard «include» Trendings «include» Text Analytics «include» «include» Brand Analysis

TABLE 28: SCENARIO SCE-IIPDD-05

Section	Description
ID	SCE-IIPDD-06
Title	Categorization
Description	Information extracted from social networks, new pages and blogs is classified based on the categories defined by Administrators
Actors	Policy Makers, End user
Objectives	Classify information into defined categories for report generation
Pre-Conditions	N/A
Process Description	1. User selects "Categorization" 2. System gathers stored information 3. <<Include *Text Analytics*>> Stored information is analysed and categorized 4. System shows information to the users: Popular categories, categorized texts… 5. User uses filters to select information to analyse 6. System recalculate information to show
Variations	
Post-Conditions	Recommendations to the user are shown based on categorization
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 29: SCENARIO SCE-IIPDD-06

Section	Description
ID	SCE-IIPDD-07
Title	Trendings
Description	Information extracted from social networks, webpages and blogs is analysed in order to discover new trends and generate new recommendations to the user based on opinion analysis
Actors	Policy Makers, End user
Objectives	Discover hidden patterns and new tendencies, makes recommendation
Pre-Conditions	N/A

Section	Description
Process Description	1. User selects “Trending” 2. System gathers stored information 3. <<Include *Text Analytics*>> Stored information is analysed (entities identification, opinion analysis) 4. System shows information to the users 5. User uses filters to select information to analyse 6. System recalculate information to show
Variations	
Post-Conditions	Recommendations to the user are shown based on discovered patterns and tendencies
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 30: SCENARIO SCE-IIPDD-07

Section	Description
ID	SCE-IIPDD-08
Title	Brand Analysis
Description	Information extracted from social networks, rss channels and blogs is analysed in order to analysed the wine market
Actors	Policy Makers, End user
Objectives	Study the market in order to create new policies or adapt the existing ones
Pre-Conditions	N/A
Process Description	1. User selects “Brand analysis” 2. System gathers stored information 3. <<Include *Text Analytics*>> Stored information is analysed 4. System shows information to the users: Popular brands, new tendencies related to the wine market. 5. User selects use filters to select information to analyse 6. System recalculate information to show
Variations	N/A
Post-Conditions	Recommendations to the user are shown based on brand analysis
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 31: SCENARIO SCE-IIPDD-08

Section	Description
ID	SCE-IIPDD-09
Title	Text Analysis
Description	Information provided by social networks, RSS channels and blogs is analysed and classified
Actors	N/A
Objectives	Opinion Analysis, Sentiment Analysis, extraction of entities
Pre-Conditions	
Process Description	1. System read the text to process 2. System analysed and categorized the text
Variations	N/A

Section	Description
Post-Conditions	Unknown at this phase, it needs to be refined in updated versions of the document
UML User Case Diagram	Could not be defined at this phase, it will be provided at the official release of the first version of the deliverable

TABLE 32: SCENARIO SCE-IIPDD-09

2.2.3 Stakeholder Requirements

The following tables contain the initial list of the stakeholder requirements for the scenarios of this use case that were described in the previous subsection.

Section	Description
ID	REQ- IIPDD -01
Title	Restricted access to configuration panel
Level of detail	Stakeholder
Type	FUNC
Description	It will be required a login and a password to configure which information has to be gathered and hierarchical categorization
Additional Information	N/A
Actor	Administrators, Policy Cloud
Priority	MAN
Reference Use Case	SCE-IIPDD-01
Success Criteria	Nobody without login/password can access to configuration panel
Expected delivery date	

TABLE 33: STAKEHOLDER REQUIREMENT REQ- IIPDD-01

Section	Description
ID	REQ- IIPDD -02
Title	Opinion Analysis
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks and news channels should be analysed in order to know people’s opinions about products & brands analysed by Appellation of origin
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success	Unknown at this phase, it needs to be refined in updated versions of the document

Section	Description
Criteria	
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 34: STAKEHOLDER REQUIREMENT REQ- IIPDD-02

Section	Description
ID	REQ- IIPDD -03
Title	Sentiment Analysis
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks and news channels should be analysed in order to know what people feels about products & brands analysed by Appellation of origin
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 35: STAKEHOLDER REQUIREMENT REQ- IIPDD-03

Section	Description
ID	REQ- IIPDD -04
Title	Hierarchical text classification
Level of detail	Stakeholder
Type	FUNC
Description	Information gathered from social networks and news channels should be classified regarding to defined hierarchical classification
Additional Information	Classification must be agreed with policy regulators and appellation of origin
Actor	Administrator, Policy Makers
Priority	MAN
Reference Use Case	SCE-IIPDD-04, SCE-IIPDD-06, SCE-IIPDD-09
Success Criteria	All texts have to be classified into one or more categories
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 36: STAKEHOLDER REQUIREMENT REQ- IIPDD-04

Section	Description
ID	REQ- IIPDD -05
Title	Extraction of entities
Level of detail	Stakeholder
Type	FUNC
Description	More relevant entities (Proper names, location, organization) should be extracted from gathered texts
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-05, SCE-IIPDD-06, SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Entities extracted from texts
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 37: STAKEHOLDER REQUIREMENT REQ- IIPDD-05

Section	Description
ID	REQ- IIPDD -06
Title	Personal Data
Level of detail	Stakeholder
Type	DATA
Description	Personal data like names, address etc. will not be stored
Additional Information	Taking into account GPDR
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 38: STAKEHOLDER REQUIREMENT REQ- IIPDD-06

Section	Description
ID	REQ- IIPDD -07
Title	Twitter information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from tweets: text, user (alias), channel, location, date), origin source. It needs to be redefined in a later phase though
Additional Information	Only information for public accounts will be collected

Section	Description
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 39: STAKEHOLDER REQUIREMENT REQ- IIPDD-07

Section	Description
ID	REQ- IIPDD -08
Title	Facebook Information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from Facebook posts: text, user (alias), channel, location, date, origin source, it needs to be redefined on a later phase though
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 40: STAKEHOLDER REQUIREMENT REQ- IIPDD-08

Section	Description
ID	REQ- IIPDD -09
Title	Instagram Information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from Instagram posts: text, user (alias), channel, location, date, origin source. It needs to be redefined in a later phase though
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	DES
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 41: STAKEHOLDER REQUIREMENT REQ- IIPDD-09

Section	Description
ID	REQ- IIPDD -10
Title	LinkedIn Information
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from LinkedIn posts: text, user (alias), channel, location, date, origin source. It needs to be redefined in a later phase though
Additional Information	Only information for public accounts will be collected
Actor	N/A
Priority	DES
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 42: STAKEHOLDER REQUIREMENT REQ- IIPDD-10

Section	Description
ID	REQ- IIPDD -11
Title	RSS & Blogs web pages
Level of detail	Stakeholder
Type	DATA
Description	Information to gather from news web pages: text, source, date, title, but it has to be refined in updated versions of the document
Additional Information	N/A
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-07, SCE-IIPDD-08, SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 43: STAKEHOLDER REQUIREMENT REQ- IIPDD-11

Section	Description
ID	REQ- IIPDD -12
Title	Information gathering in near real-time
Level of detail	Stakeholder
Type	PERF
Description	Data from different sources should be taken in near real-time
Additional Information	Every 1h but it might need to be refined in updated versions of the document

Section	Description
Actor	N/A
Priority	MAN
Reference Use Case	N/A
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 44: STAKEHOLDER REQUIREMENT REQ- IIPDD-12

Section	Description
ID	REQ- IIPDD -13
Title	Data Analysis in near real-time
Level of detail	Stakeholder
Type	PERF
Description	Information collected from different dataset should be analysed immediately
Additional Information	Just after data is collected, but it might need to be refined in updated versions of the document
Actor	N/A
Priority	MAN
Reference Use Case	SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 45: STAKEHOLDER REQUIREMENT REQ- IIPDD-13

Section	Description
ID	REQ- IIPDD -14
Title	Dashboard displaying
Level of detail	Stakeholder
Type	L&F
Description	Dashboard should show more relevant information at a glance
Additional Information	N/A
Actor	End User, Policy Maker
Priority	DES
Reference Use Case	SCE-IIPDD-05, SCE-IIPDD-06, SCE-IIPDD-07, SCE-IIPDD-08
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 46: STAKEHOLDER REQUIREMENT REQ- IIPDD-14

Section	Description
ID	REQ- IIPDD -15
Title	Web design
Level of detail	Stakeholder
Type	L & F
Description	Information should be displayed on a web page
Additional Information	N/A
Actor	
Priority	MAN
Reference Use Case	SCE-IIPDD-01, SCE-IIPDD-05, SCE-IIPDD-06, SCE-IIPDD-07, SCE-IIPDD-08
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 47: STAKEHOLDER REQUIREMENT REQ- IIPDD-15

Section	Description
ID	REQ- IIPDD -16
Title	AA Standars
Level of detail	Stakeholder
Type	USE
Description	Web page should accomplish AA standards of accessibility
Additional Information	N/A
Actor	End User, Policy makers
Priority	DES
Reference Use Case	SCE-IIPDD-01, SCE-IIPDD-05, SCE-IIPDD-06, SCE-IIPDD-07, SCE-IIPDD-08
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 48: STAKEHOLDER REQUIREMENT REQ- IIPDD-16

Section	Description
ID	REQ- IIPDD -17
Title	Mobile platforms
Level of detail	Stakeholder
Type	USE
Description	Designed platform should be accessible through mobile devices
Additional Information	N/A
Actor	End User, Policy makers

Section	Description
Priority	DES
Reference Use Case	SCE-IIPDD-01, SCE-IIPDD-05, SCE-IIPDD-06, SCE-IIPDD-07, SCE-IIPDD-08
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 49: STAKEHOLDER REQUIREMENT REQ- IIPDD-17

Section	Description
ID	REQ- IIPDD -18
Title	Working hours
Level of detail	Stakeholder
Type	SUP
Description	Gathering & analysis information should be working 24x7 Web page should be available 24x7
Additional Information	N/A
Actor	N/A
Priority	DES
Reference Use Case	SCE-IIPDD-09
Success Criteria	Unknown at this phase, it needs to be refined in updated versions of the document
Expected delivery date	Unknown at this phase, it needs to be refined in updated versions of the document

TABLE 50: STAKEHOLDER REQUIREMENT REQ- IIPDD-18

2.3 UC#3: Facilitating urban policy making and monitoring through crowdsourcing data analysis

2.3.1 Goals and Objectives

The aim of this use case is to support Sofia Municipality's policy making in important areas of citizen's areas of everyday life. By improving the policy making in these areas, the overall quality of citizen's life will be improved, which is the overall goal of this project.

The use case objectives are through PolicyCLOUD big data streaming and real-time big data platform to improve operational efficiency, transparency, and decision making. The PolicyCLOUD visualisation technologies will enable policy makers to identify issues, trends, and policy effects and interactions. The PolicyCLOUD analytics technologies will enable to discover insights and find meaningful explanations about the effects of policies.

The described scenarios include main policy making areas, aimed at improving under PolicyCLOUD project, namely:

- transport
- parking
- road infrastructure
- waste collection and waste distribution
- air quality

2.3.2 Description of Scenarios

Sofia municipality is constantly working to improve the urban environment and meet the challenges that the city facing. Evidence-based policy making is crucial for addressing urban challenges in a cost-efficient way, however there is yet no established process to incorporate data into policy making. The PolicyCLOUD project will support Sofia municipality to address this challenge by adapting the design of its policies, considering analytics' results that combine information (big data) of sectors, related to a) transport, parking and road infrastructure; b) waste collection and waste disposal; c) cleanliness of public spaces; d) ecology, green systems; e) violation of public order; and others, of importance to citizens. Proposed technologies will offer the advantage to interpret, manage and analyse big amounts of data both from existing data sources (citizens' Contact Centre) and from new data sources (open data sets that will become available).

The use case scenario will address policies in the mentioned domains, where PolicyCLOUD big data analytics technologies can help in identifying hazardous factors and risk situations for citizens.

The source of the data is the citizens' Contact Centre, which is operational since 2014 and it is unique point of direct communication with citizens, industries and institutions to report non-urgent alerts on deviations from normal urban environment. Citizens can file signals for waste collection and disposal, road and traffic problems, general public infrastructure, ecology, public spaces (playgrounds, public gardens and parks) etc.

By using the powerful tools provided by PolicyCloud project, Sofia Municipality will be able to carry out a detailed analysis of the territorial distribution of the signals by categories / types, areas, districts, major transport roads, etc. The results of the analysis will allow the municipal and district administrations to identify the problems in the urban environment and to adopt or modify adequate policy making decisions on budget planning and effective use of budget and public resources. It will also help Sofia Municipality be focused on improving its policy making, related to better control and monitoring in these sectors, as well as preventing/ avoiding risky or conflicting situations from happening.

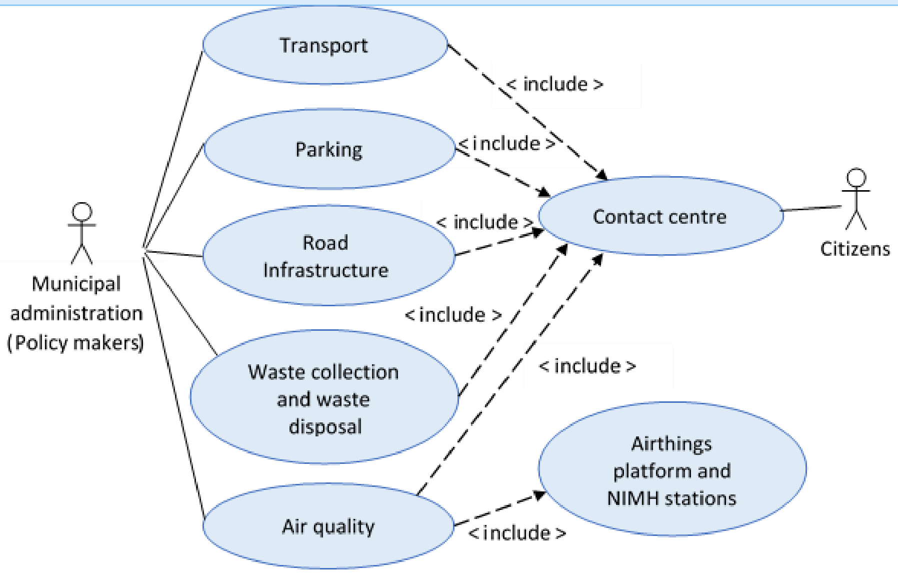
Section	Description
ID	SCE-RLIMP-01
Title	Transport
Description	Transport of Sofia is a very complex network. New business and urban areas are being constantly developed and adjustments to the current transport schemes are done. There are 200 smart crossroads equipped with cameras and sensors. The city Centre of Traffic Control is monitoring and managing the data to ensure no delays in public transportation. Periodically, to tackle air pollution problem and incentivize citizens to use the public transport, Sofia Municipality is introducing daily "Green Ticket" for a substantially lower daily price of EUR 0.5 for the whole network, other actions include offering free parking slots in the buffer parkings to stimulate the use of the metro as a means of transport. Sofia Municipality is gradually renovating the vehicle fleet. By using signals from Contact Centre (and potentially data from the Centre of Traffic Control and the Urban Mobility Centre) potential structural changes and improvement of Sofia Municipality transport policies can be analysed before implementation as well as current policies can be assessed.
Actors	Urban Mobility Centre (Municipal enterprise), Metropoliten EAD, Centre of Traffic Contol, Municipal administration, citizens, PolicyCLOUD
Objectives	improve quality of service improve transport times and better connections for citizens assess multimodal pricing schemes and initiatives such as "Green ticket"
Pre-Conditions	• Data is available from contact centre • Potentially data from Traffic Control Centre and the Centre for Urban Mobility can be obtained (tbc)
Process Description	• Municipality submits collected data for analysis • PolicyCLOUD technological providers analyses data • Municipality gets results from analyses • Municipality improved policy making
Variations	N/A
Post-Conditions	objectives completed
UML User Case Diagram	 Transport < include > Parking <include > < include > Road Infrastructure Contact centre Citizens Municipal administration (Policy makers) <include > < include > Waste collection and waste disposal Airthings platform and NIMH stations Air quality < include >

TABLE 51: SCENARIO SCE-RLIMP-01

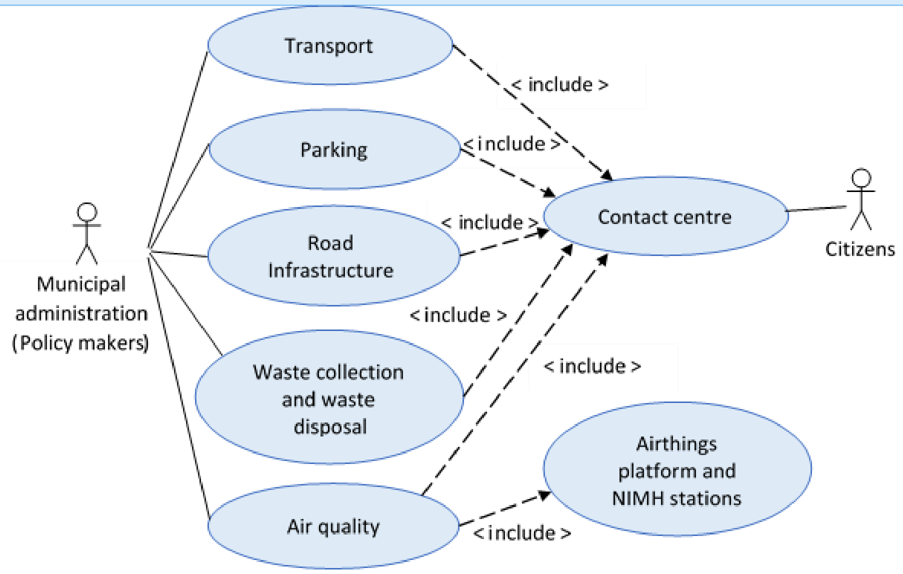
Section	Description
ID	SCE-RLIMP-02
Title	Parking
Description	Parking in Sofia is a complex issue with multiple providers. There are a number of municipal parking lots and municipal paid zones within the centre and many private providers. The aim is to optimize provision of parking services for the citizens.
Actors	Urban Mobility Centre (Municipal enterprise), Municipal administration, private providers citizens, PolicyCloud
Objectives	• Adopt quantity measures for better parking management • Improve overall parking capabilities
Pre-Conditions	• Data is available from contact centre • Potentially data from municipal parking lots can be obtained (tbc)
Process Description	• Municipality submits collected data for analysis • PolicyCLOUD technological providers analyses data • Municipality gets results from analyses • Municipality improved policy making
Variations	N/A
Post-Conditions	Objectives completed
UML User Case Diagram	 Transport < include > Parking <include > < include > Contact centre Citizens Road Infrastructure Municipal administration (Policy makers) < include > Waste collection and waste disposal < include > Airthings platform and NIMH stations Air quality < include >

TABLE 52: SCENARIO SCE-RLIMP-02

Section	Description
ID	SCE-RLIMP-03
Title	Road Infrastructure
Description	Road infrastructure is one of the most important and budget consuming element from the urban environment, that impacts citizens' everyday life. Reliable analysis is needed on current situation in all 24 district administrations, in order to foresee and improve long term policy making in the area of read infrastructure.
Actors	District administrations, Municipal administration, citizens, municipal road companies; Investors and businesses, PolicyCLOUD
Objectives	• Improving long term policy making in the area of road infrastructure

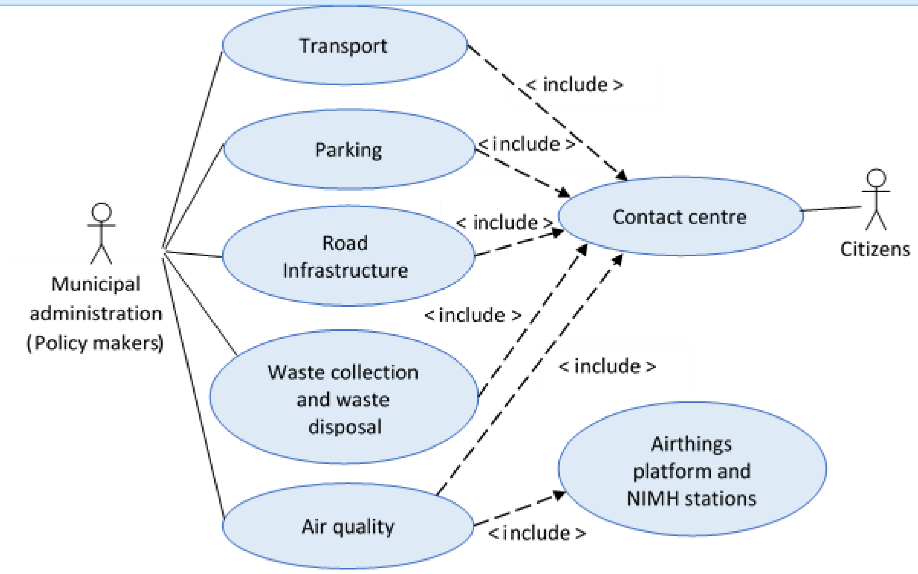
Section	Description
	• Better envisioning and capacity building of district administrations and municipal administration in solving road infrastructure problems
Pre-Conditions	• Contact centre data • Other municipal data tbc.
Process Description	• Municipality submits collected data for analysis • PolicyCLOUD technological providers analyses data • Municipality gets results from analyses • Municipality improved policy making
Variations	N/A
Post-Conditions	Objectives completed
UML User Case Diagram	 Transport Parking Road Infrastructure Waste collection and waste disposal Air quality Contact centre Airthings platform and NIMH stations Municipal administration (Policy makers) Citizens < include > < include > < include > < include > < include > < include >

TABLE 53: SCENARIO SCE-RLIMP-03

Section	Description
ID	SCE-RLIMP-04
Title	Waste collection and waste disposal
Description	Sofia Municipality collects information about waste collection, using smart meters like smart bins, smart garbage trucks, etc. Gathering and analyzing large amounts of data, related to waste management will help Sofia Municipality improve its policy making in the area of urban sustainability and will help the city to become a greener city.
Actors	District administrations, Municipal administration, citizens, waste collection companies, Sofia recycling plant, PolicyCloud
Objectives	• More efficient way of waste collection • Improvement of long term planning and policy making of waste collection and waste disposal using smart meters
Pre-Conditions	• Contact centre data • Potentially data available from smart metering in waste bins and trucks (tbc)
Process Description	• Municipality submits collected data for analysis • PolicyCLOUD technological providers analyses data • Municipality gets results from analyses

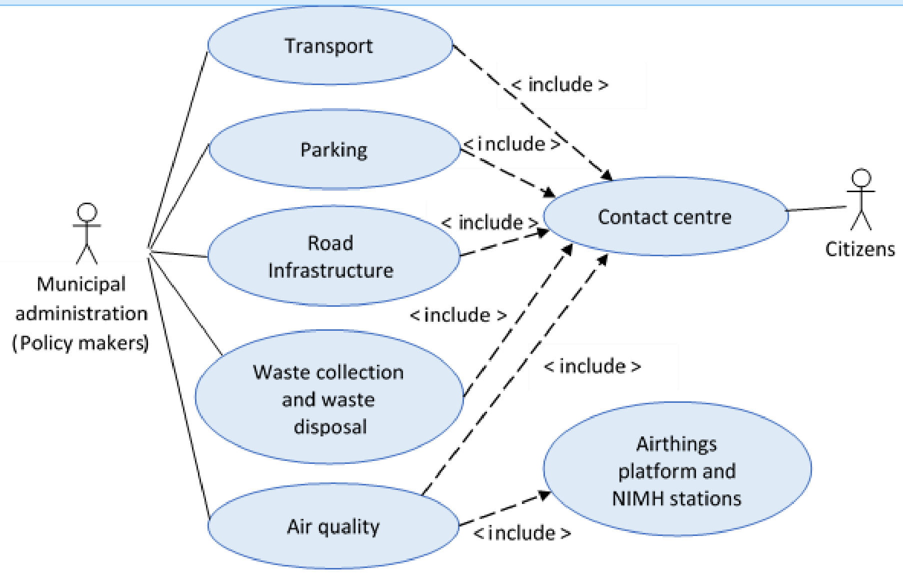
Section	Description
	• Municipality improved policy making
Variations	N/A
Post-Conditions	Objectives completed
UML User Case Diagram	 Transport < include > Parking < include > Road Infrastructure < include > Contact centre Citizens Municipal administration (Policy makers) < include > Waste collection and waste disposal < include > Airthings platform and NIMH stations Air quality < include >

TABLE 54: SCENARIO SCE-RLIMP-04

Section	Description
ID	SCE-RLIMP-05
Title	Air quality
Description	Under the Airthings project Sofia built a local network of IoT sensors for monitoring and measurement of the air quality, a smart Internet platform was developed to visualize and store air data. It uses cloud technologies combined with analytical functionality and advanced machine learning opportunities to enable the city administration to take timely action to introduce actions aimed at improving the air quality. The system interface provides the general public with visualization and machine-readable data available through the web and applications notifications for smartphone and tablets. This project is additional to an earlier set up by the National Institute for Meteorology and Hydrology (NIMH) developed for Sofia Municipality, which provided a system that warns about the danger of high levels of air pollution 48 hours earlier. Executive Environment Agency) helps Sofia Municipality improve its long term policy making and taking adequate short-term decisions, concerning air quality.
Actors	Municipal administration, citizens, PolicyCLOUD
Objectives	• Improvement of long-term policy making in the area of air quality
Pre-Conditions	• Data available from contact centre, and the Airthings platform and the NIMH stations;
Process Description	• Municipality submits collected data for analysis • PolicyCLOUD technological providers analyses data • Municipality gets results from analyses • Municipality improved policy making

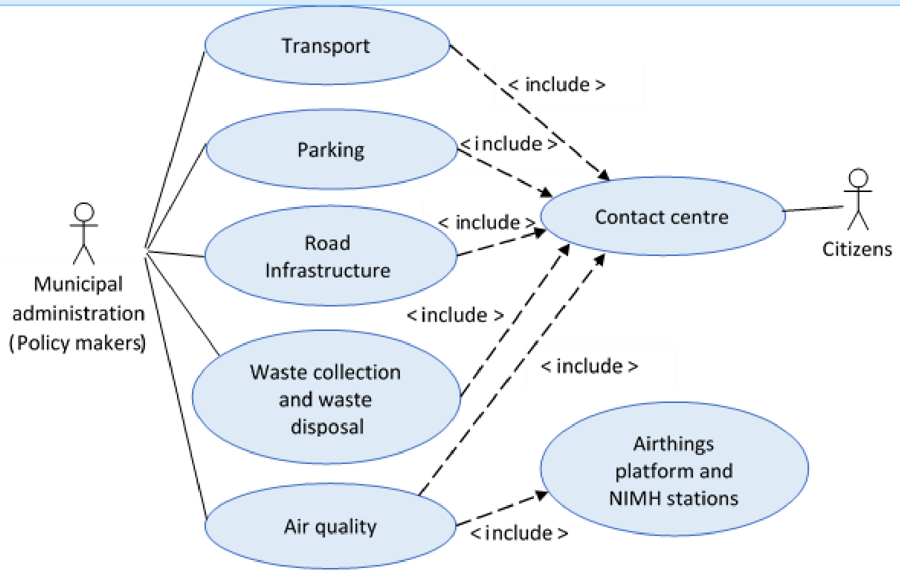
Section	Description
Variations	N/A
Post-Conditions	Objectives completed
UML User Case Diagram	 Transport Parking Road Infrastructure Waste collection and waste disposal Air quality Municipal administration (Policy makers) Contact centre Citizens Airthings platform and NIMH stations < include > < include > < include > < include > < include > < include >

TABLE 55: SCENARIO SCE-RLIMP-05

2.3.3 Stakeholder Requirements

The following tables contain the initial list of the stakeholder requirements for the scenarios of this use case that were described in the previous subsection.

Section	Description
ID	REQ- RLIMP-01
Title	Provide assessment and visualization
Level of detail	Stakeholder
Type	FUNC
Description	For any policy scenario, the system should be able to predict a set of outcomes.
Additional Information	N/A
Actor	Sofia Municipality
Priority	MAN (mandatory requirement) for the analysis DES (desirable requirement) for visualization
Reference Use Case	SCE-RLIMP-01, SCE-RLIMP-02, SCE-RLIMP-03, SCE-RLIMP-04, SCE-RLIMP-05
Success Criteria	Unknown at this time
Expected delivery date	Unknown at this time

TABLE 56: STAKEHOLDER REQUIREMENT REQ- RLIMP-01

2.4 UC#4: Predictive analysis towards unemployment risks identification and policy making

2.4.1 Goals and Objectives

The goal of this use case is to assist policy makers in creating effective policies that will address employment figures. The overall goal of this is use case is for Policy makers to be able to use statistics from predictive algorithms from the toolkit to assist in making decision during policy creation process. The main objective will be to design the algorithms that will help predict future trends using the provided unemployment database.

2.4.2 Description of Scenarios

Due to the spreading of the pandemic disease in the municipality of London, this pilot case could not proceed adequately at the first semester, as the main focus was on the confrontation of COVID-19. Due to this, the description of the use case scenario was postponed to be delivered on the next iteration of this document.

2.4.3 Stakeholder Requirements

The following tables contain the initial list of the stakeholder requirements for the scenarios of this use case that were described in the previous subsection.

Section	Description
ID	REQ- PAUNRI-01
Title	Analysis capabilities
Level of detail	Stakeholder
Type	ENV
Description	The PolicyCLOUD toolkit should be able to produce some form of visualisation or report that can help create policies
Additional Information	Additional information that might be need for this requirement
Actor	End User, Policy makers
Priority	MAN
Reference Use Case	UC#4
Success Criteria	The success of this requirement can be tracked based from the feedback from users that have completed questionnaires. The questionnaire will include clear questions that will indicate whether the user experience with the platform was positive or negative.
Expected delivery date	

TABLE 57: STAKEHOLDER REQUIREMENT REQ - PAUNRI-01

Section	Description
ID	REQ- PAUNRI-02
Title	User interface
Level of detail	Stakeholder
Type	L&F
Description	The description of the requirement
Additional Information	
Actor	The list of different actors that are related to this requirement
Priority	DES
Reference Use Case	UC#4
Success Criteria	Involved stake the success of this requirement can be tracked based from the feedback from users that have completed questionnaires. The questionnaire will include clear questions that will indicate whether the user experience with the platform was positive or negative.
Expected delivery date	

TABLE 58: STAKEHOLDER REQUIREMENT REQ - PAUNRI-02

Section	Description
ID	REQ- PAUNRI-03
Title	Secure infrastructure
Level of detail	Stakeholder
Type	ENV
Description	The description of the requirement
Additional Information	
Actor	End User, Policy makers
Priority	DES
Reference Use Case	UC#4
Success Criteria	
Expected delivery date	

TABLE 59: STAKEHOLDER REQUIREMENT REQ - PAUNRI-03

Section	Description
ID	REQ- PAUNRI-04
Title	Help page/ Support documentation
Level of detail	Stakeholder
Type	SUP
Description	The description of the requirement
Additional Information	
Actor	End User, Policy makers
Priority	ENH
Reference Use Case	UC#4
Success Criteria	User should be able to access documentation or a user help page that will assist and display explanation for each of the toolkits action buttons.
Expected delivery date	

TABLE 60: STAKEHOLDER REQUIREMENT REQ - PAUNRI-04

3 Use case datasets and data regulatory constraints

This section contains the definition of the all available datasets that will be used in the scope of the PolicyCLOUD project, along with potential data regulatory constraints that might be needed to be enforced when these datasets are being accessed by third parties or are being collected and stored in a cloud environment, such as the deployment of the PolicyCLOUD platform outside of the proprietary's premises. The list of these requirements for data management and regulatory constraints will mainly drive the implementation of the corresponding components of the overall PolicyCLOUD architecture, mainly the data repository, the data fusion and the definition of the data governance model and the protection and privacy enforcement. These requirements are listed in the following two subsections and are presented per use cade.

3.1 Dataset Specifications

3.1.1 UC#1: Participatory policies against radicalization

Section	Description
ID	DS-PPR-01
Title	Twitter
Description	Relevant posts published by users
Owner	Twitter
Licence/Privacy	Twitter license, to be discussed
Data type	Text and images
Type of Process (Stream or Static data)	streaming
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every minute
Documentation	Twitter documentation

TABLE 61: DATASET REQUIREMENT DS -PPR-01

Section	Description
ID	DS-PPR-02
Title	Facebook
Description	Relevant posts and comments published in open groups
Owner	Facebook
Licence/Privacy	Facebook license, to be discussed
Data type	Text and images

Section	Description
Type of Process (Stream or Static data)	streaming
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every minute
Documentation	Twitter documentation

TABLE 62: DATASET REQUIREMENT DS-PPR-02

Section	Description
ID	DS-PPR-03
Title	Reddit
Description	Relevant posts published by users
Owner	n/a
Licence/Privacy	n/a
Data type	Text, images
Type of Process (Stream or Static data)	Streaming
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every minute
Documentation	Reddit documentation

TABLE 63: DATASET REQUIREMENT FOR DS-PPR-03

Section	Description
ID	DS-PPR-04
Title	RSS Feeds
Description	Relevant news available on the web
Owner	It is not sure at the moment
Licence/Privacy	It is not sure at the moment
Data type	unstructured (i.e. text, article, image)
Type of Process (Stream or Static data)	stored on data repository

Section	Description
Data Format	HTML
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every hour
Documentation	

TABLE 64: DATASET REQUIREMENT FOR DS-PPR-04

Section	Description
ID	DS-PPR-05
Title	Global Terrorism Database (https://www.start.umd.edu/gtd/)
Description	open-source database including information on domestic and international terrorist attacks around the world from 1970 and includes more than 190,000 cases.
Owner	The National Consortium for the Study of Terrorism and Responses to Terrorism (START)
Licence/Privacy	Open source for research purposes, Licenses for commercial purposes
Data type	Structured (i.e. text, article, image)
Type of Process (Stream or Static data)	stored on data repository
Data Format	Structured data (.xlsx)
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every week
Documentation	Not yet clear

TABLE 65: DATASET REQUIREMENT FOR DS-PPR-05

3.1.2 UC#2: Intelligent policies for the development of denomination of origin

In the following tables, the datasets that are planned to be used by the use case of the *Intelligent policies for the development of denomination of origin* are being included.

Section	Description
ID	DS-IIPDD-01
Title	CAP
Description	The Common Agricultural Policy (CAP) is the agricultural policy of the European Union. It implements a system of agricultural subsidies and other programmes. It was introduced in 1962 and has undergone several changes since then to reduce the cost (from 73% of the EU budget in 1985 to 37% in 2017) and to also consider rural development in its aims. It

Section	Description
	has been criticised on the grounds of its cost, and its environmental and humanitarian impacts.
Owner	Open Data Aragon (https://opendata.aragon.es/)
Licence/Privacy	It can be imported in a PolicyCLOUD but it needs to be reviewed in the forthcoming versions of this deliverable
Data type	Semi-structural
Type of Process (Stream or Static data)	Database
Data Format	Virtuoso (triplets), JSON or XML
Data Store	Virtuoso
Recommended API	Sparql
Data Volume	Millions but we require few of them
Data Velocity	This is not yet clear at this phase. It will refined in the updated versions of the document
Documentation	https://opendata.aragon.es/datos/catalogo?texto=pac

TABLE 66: DATASET REQUIREMENT FOR DS-IIPDD-01

Section	Description
ID	DS-IIPDD-02
Title	Wine register
Description	Sigpac reference, variety, cultivation year, area (hec)
Owner	Aragon Government
Licence/Privacy	It can be imported in a PolicyCLOUD, but it needs to be reviewed in the forthcoming versions of this deliverable
Data type	Structural
Type of Process (Stream or Static data)	static
Data Format	csv
Data Store	Web links to files
Recommended API	REST API
Data Volume	1 Gbytes
Data Velocity	This is not yet clear at this phase. It will be refined in the updated versions of the document
Documentation	https://www.aragon.es/en/-/consultas-sigpac

TABLE 67: DATASET REQUIREMENT FOR DS-IIPDD-02

Section	Description
ID	DS-IIPDD-03
Title	Production data
Description	Data production per grape variety

Section	Description
Owner	SARGA
Licence/Privacy	It can be imported in a PolicyCLOUD, but it needs to be reviewed in the forthcoming versions of this deliverable
Data type	Structural
Type of Process (Stream or Static data)	Static data
Data Format	Tables ODBC
Data Store	Microsoft SQL
Recommended API	JDBC
Data Volume	2 GBytes
Data Velocity	This is not yet clear at this phase. It will be refined in the updated versions of the document
Documentation	To be described

TABLE 68: DATASET REQUIREMENT FOR DS-IIPDD-03

Section	Description
ID	DS-IIPDD-04
Title	Twitter data
Description	Information provided by users about wine varieties, brands
Owner	N/A
Licence/Privacy	Twitter license to be discussed
Data type	Text and images
Type of Process (Stream or Static data)	streaming
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	1 Terabyte for this scenario
Data Velocity	Every hour 10 minutes
Documentation	Twitter documentation

TABLE 69: DATASET REQUIREMENT FOR DS-IIPDD-04

Section	Description
ID	DS-IIPDD-05
Title	Facebook data
Description	Information provided by users about wine varieties, brands
Owner	To be studied
Licence/Privacy	Facebook license to be discussed

Section	Description
Data type	Text, images
Type of Process (Stream or Static data)	Streaming
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	1 Terabyte
Data Velocity	Every 10 minutes
Documentation	Facebook doc

TABLE 70: DATASET REQUIREMENT FOR DS-IIPDD-05

Section	Description
ID	DS-IIPDD-06
Title	Instagram
Description	Information provided by users about wine varieties, brands
Owner	N/A
Licence/Privacy	Instagram license
Data type	unstructured (i.e. text, article, image)
Type of Process (Stream or Static data)	Streaming and stored on data repository
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	1 Terabyte
Data Velocity	Every 10 minutes
Documentation	Instagram

TABLE 71: DATASET REQUIREMENT FOR DS-IIPDD-06

Section	Description
ID	DS-IIPDD-07
Title	LinkedIn
Description	Information provided by users about wine varieties, brands
Owner	N/A
Licence/Privacy	LinkedIn license to be discussed
Data type	unstructured (i.e. text, article, image)
Type of Process (Stream or Static data)	Streaming and stored on data repository

Section	Description
Data Format	JSON
Data Store	N/A
Recommended API	REST API
Data Volume	1 Terabyte
Data Velocity	Every 10 minutes
Documentation	

TABLE 72: DATASET REQUIREMENT FOR DS-IIPDD-07

Section	Description
ID	DS-IIPDD-08
Title	News Webpages /blogs
Description	News about wine, brands, production of wine
Owner	NewsPapers
Licence/Privacy	To be reviewed
Data type	unstructured (i.e. text, article, image)
Type of Process (Stream or Static data)	stored on data repository
Data Format	HTML
Data Store	N/A
Recommended API	REST API
Data Volume	This is not yet clear at this phase. It will be refined in the updated versions of the document
Data Velocity	Every hour
Documentation	

TABLE 73: DATASET REQUIREMENT FOR DS-IIPDD-08

3.1.3 UC#3: Facilitating urban policy making and monitoring through crowdsourcing data analysis

In the following table the datasets that are planned to be used by the use case of the *Facilitating urban policy making and monitoring through crowdsourcing data analysis* are being included.

Section	Description
ID	DS-RLIMP-01
Title	SofiaMunicipalitySignals
Description	Signals from citizens, coming through the contact centre of the municipality. For each signal submitted, the following data is contained: • signal ID • signal category; • signal type; • day, month, year, date and time of reporting;

Section	Description
	- signal location geographic coordinates - address (text field) - short signal description (text field) - photos - detailed signal description (text field) - signal status - answers to clarifying questions, related to the signal. In addition, per signal there is the following information: - status (waiting, being processed, resolved) with time stamps - deadline to process - resolution details
Owner	The data is owned by Sofia Municipality. A limited number of signals (the last 20) are displayed on the platform
Licence/Privacy	The data can be imported in a PolicyCLOUD environment.
Data type	The data is structural – XML
Type of Process (Stream or Static data)	The data is static. However, the status and resolution of signals changes and can be updated periodically. It can also be updated with additional records.
Data Format	XML file
Data Store	MSSQL 2014 Database
Recommended API	Generating XML file with the necessary data directly or through REST API
Data Volume	Depends on the period taken. Data per month 5000 * 70 KB/ month = 350 MB/ month This includes photo images data, without photo images volumes would be ten times lower. However, this monthly data is estimated based on the signals, coming through the web-based system only. In April 2020 Sofia Municipality is launching mobile app in addition to the web-based system, so volume of signals is expected to increase.
Data Velocity	5000 signals per month However, this monthly data is estimated based on the signals, coming through the web-based system only. In April 2020 Sofia Municipality is launching mobile app in addition to the web-based system, so volume of signals is expected to increase.
Documentation	The data is not public. A limited number of signals (the latest 20) can be seen on https://call.sofia.bg/

TABLE 74: DATASET REQUIREMENT FOR DS-RLIMP-01

3.1.4 UC#4: Predictive analysis towards unemployment risks identification and policy making

The following dataset is planned to be used in the scope of this pilot.

Section	Description
ID	DS-PAUNRI-01
Title	Unemployment Claimant Count LATEST
Description	A short description of the content of the dataset
Owner	This is an open and publicly available dataset

Section	Description
Licence/Privacy	State the type of license and privacy considerations. Under data owner authorisation the dataset can be imported into a PolicyCLOUD environment
Data type	Structural (CSV)
Type of Process (Stream or Static data)	This dataset will be stream data that will be periodically updated
Data Format	The data will be stored as CSV
Data Store	N/A
Recommended API	The recommended API for this dataset is the REST API
Data Volume	Estimation of the volume both in terms of storage size and number of records
Data Velocity	N/A
Documentation	https://opendata.camden.gov.uk/Business-Economy/Unemployment-Claimant-Count-LATEST/g3p6-usd3 - This should be documented where

TABLE 75: DATASET REQUIREMENT FOR DS-PAUNRI-01

3.2 Regulatory Constraint Requirements

This section will contain the requirements regarding the regulatory constraints of each pilot. However, at this phase of the project it is very early for this type of analysis due to the fact that it requires the specification of the involved datasets of each use cases that was reported at the previous subsection. As this work was delivered at the end of M06 and reported at this version of the delivery, the analysis of the regulatory constraint requirements has not been started yet. The results of this analysis are planned to be included in the second version of this deliverable.

4 Platform Roles

The following table contains the list of all different roles of actors that are related with the development, deployment, operation and usage of all solutions that are offered by the PolicyCLOUD platform, along with a description of these roles. These roles are involved in different system and software requirements, as will be further analysed in the corresponding sections 1 and 0.

The roles listed in the table are non-exhaustive and need to be further extended (if needed) by the relevant partners that are domain experts in the field of PolicyCLOUD, as the project is progressing.

ID	Name	Description
ROL-01	Data Owner	PolicyCLOUD provides a cloud Gateway component that can be used in order to push data coming from different data owners in various formats that can be either static or streaming data, which is finally stored in the data store of the platform and is accessible by its analytical tools.
ROL-02	Data Engineers	PolicyCLOUD offers via the Data Analytics component a framework to register their analytical tools that can use the common data repository of the project in order to perform analytical tasks on the stored data. It can also rely on the intermediate results of other tools and feed with them her model.
ROL-03	Policy Makers	PolicyCLOUD offers the Policy Development Toolkit that allows the policy makers to create and evaluate new policies in different domains, associate the policies with specific KPIs and validate them by triggering the execution of one of the PolicyCLOUD's analytical tools by seeing the results visualized in the toolkit.
ROL-04	Data Scientist	PolicyCLOUD offers to the data scientist the Data Marketplace in order to explore and validate the provided analytical tools with different target datasets and experimenting with extended datasets.

TABLE 76: PLATFORM ROLES

5 System Requirements

This section will present the system requirements of PolicyCLOUD, which represent the technical specifications for the platform at the systemic level. They defined the services along with the functionalities and their interfaces of the major building blocks that formulate the overall platform and need to meet the stakeholder requirements, as defined in section 1. They answer to what characteristics the system needs to possess and to what degree to satisfy the stakeholder requirements.

The system requirements that are defined in this section are organized per major building block that corresponds to a specific capability. All these capabilities are depicted in the full stack that is presented in Figure 2. As it can be depicted from the figure, PolicyCLOUD platform proves a full stack of capabilities that aim for the data acquisition and collection via its gateways, the persistent store of data in the data repository, the deployment of numerous and heterogeneous analytical tools that can make use of this data, the incorporation of reusable models exploited by these tools, and finally, the creation and evaluation of policies via the relevant policy development toolkit, that relies on the use of the models and tools of the underlying layer, and the experimentation with additional datasets that can be discovered and exploitable via the use of the Marketplace.

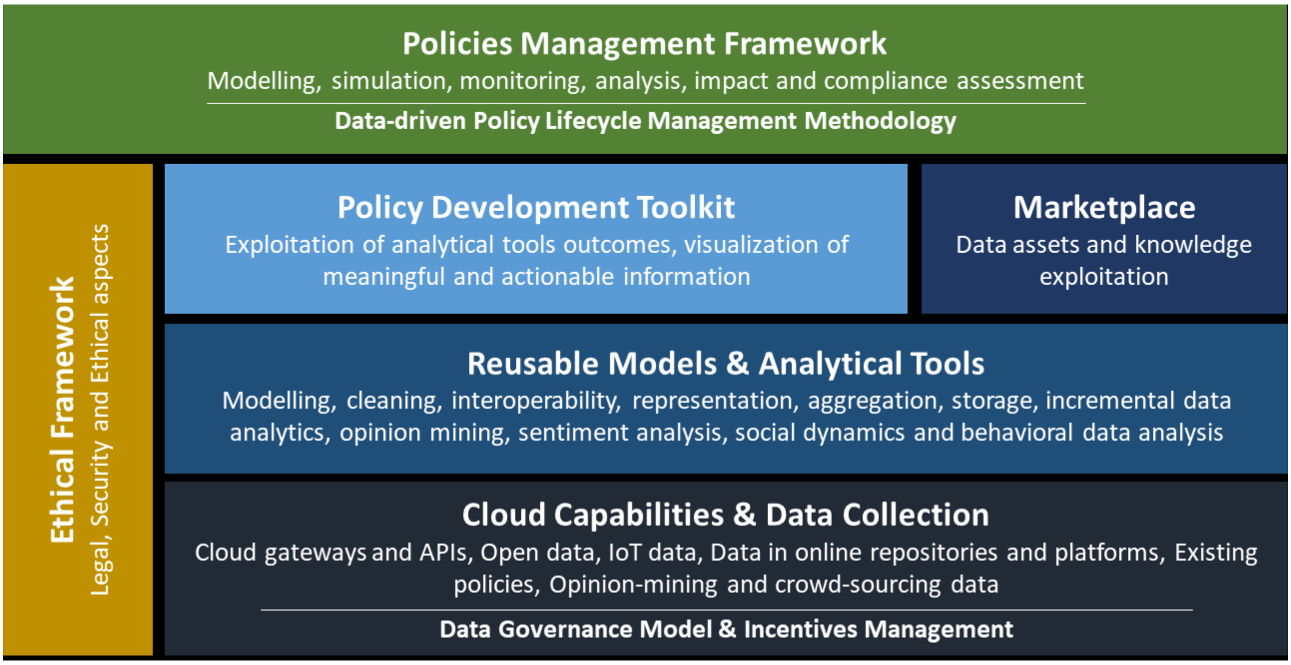

Policies Management Framework
Modelling, simulation, monitoring, analysis, impact and compliance assessment
Data-driven Policy Lifecycle Management Methodology
Ethical Framework
Legal, Security and Ethical aspects
Policy Development Toolkit
Exploitation of analytical tools outcomes, visualization of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

FIGURE 2: POLICYCLOUD CORE SYSTEM CAPABILITIES

In more detail, PolicyCLOUD consists of the following different capability categories:

- **Cloud Capabilities and Data Collection**: This layer consists of all operations required for cloud registration and resource provisioning in order for all the platform components and analytical tools to be instantiated and deployed. It additionally consists of various mechanisms for data acquisition, collection, cleaning and persistent storage, while also providing the means for a seamless manner for data retrieval of data that might exist inside the platform, or that can be accessed in cases a dataset must stay on-premise and cannot be loaded to the platform.
- **Data Governance Model and Incentives Management**: This layer is taking care of the various data regulatory constraints that are imposed by the use cases and the owners of the data, in order to allow them access to specific data resources or not. It is used by the data collection mechanisms in order to

allow for the data acquisition, while on the same time, it is used by the seamless data retrieval component of the polyglot data store in order to allow the establishment of a data connection to external resources for data retrieval.

- **Reusable Models and Analytical Tools**: This layer contains all various analytical tools that are natively supported and implemented by the PolicyCLOUD platform, such as opinion mining and sentiment analysis, behavioural data analytics, situational knowledge acquisition, social dynamics etc. It also contains the reusable models that can be shared by all tools, thus increasing the level of interoperability of all the deployed tools.
- **Policy Development Toolkit**: This component offers a framework for policy makers and domain experts to define implement and evaluate a new policy that is related with specific KPIs and trigger the execution of an analytical task in order to validate the results against the already defined KPIs. It has access to the underlying analytical tools provided by the underlying layer and can make use of intermediate or previous produced results to validate the KPIs in the process of the time, making use of the data stored persistently in the PolicyCLOUD or data that can be accessed via the use of the polyglot capabilities of the data repository of the platform.
- **Data Marketplace**: This component of PolicyCLOUD provides to the data analysts and policy experts the ability to discover new available datasets that can be offered via the platform, and either trigger the data acquisition process in order to ingest the corresponding datasets into the data repository, so that they can be available by the analytical tools, or to provide the capability to connect remotely to the external data source, if this is allowed by the data regulatory constraints that are validated by the Data Governance Model.
- **Ethical Framework**: At this phase of the project, there has not been enough progress of the exact specification of the Ethical Framework and more information has been planned to be included in the next version of the deliverable.
- **Policy Management Framework**: At this phase of the project, there has not been enough progress of the exact specification of the Policy Management Framework and more information has been planned to be included in the next version of the deliverable.

The following subsections contain the list of the system requirements that are imposed by each of the aforementioned system capabilities of the PolicyCLOUD platform. Due to the early state of the work that is being done in parallel in T2.2 which is responsible for the definition of the overall design of the architecture of the platform, these system building blocks are not yet finalized, and not all system requirements have been identified yet. Due to this, the refined list of the system requirements will be included in the second version of the document that is planned to be delivered on M12, while the complete list of all system requirements will be delivered in the final version of this deliverable that is planned to be released on M22, which will drive the finalization of the architecture of the PolicyCLOUD platform.

5.1 Cloud Capabilities and Data Collection

Section	Description
ID	REQ- SY-CCDC-01
Title	Cloud data storage should scale out in order to store data whose size is getting increased
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	In cases that a use case is increasing the volume of the data to be stored, the cloud data storage should be able to scale out in order that all dataset can be persistently stored
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-01, ROL-02
Source	N/A
Success Criteria	Stress the storage to reach its limits and check that it does not fail
Expected delivery date	M36 (3rd version of prototypes)

TABLE 77: SYSTEM REQUIREMENT REQ -SY-CCDC-01

5.2 Data Governance Model and Incentives Management

Section	Description
ID	REQ- SY-DGMIM-01
Title	Incentive Management Visualizations
Level of detail	System
Type	FUNC (function)
Description	The Incentives Management component(s) will interact with the Policy Maker and with the participants through the PDT.
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	all
Role	ROL-01, ROL-03
Source	PDT Technology
Success Criteria	N/A
Expected delivery date	M24 (2nd version of prototypes)

TABLE 78: SYSTEM REQUIREMENT REQ - SY-DGMIM-01

Section	Description
ID	REQ- SY-DGMIM-02
Title	Incentive Management REST API
Level of detail	System
Type	FUNC (function)
Description	The component(s) provided in the context of the Incentives Management task will expose their features through an interface in the form of a REST API.
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	all
Role	ROL-02
Source	REST technology
Success Criteria	N/A
Expected delivery date	M12 (1st version of prototypes)

TABLE 79: SYSTEM REQUIREMENT REQ- SY-DGMIM-02

Section	Description
ID	REQ- SY-DGMIM-03
Title	Incentive Management Storage Backend
Level of detail	System
Type	DATA (data)
Description	The domain objects resulting from the Incentives Management component(s) will be stored in the PolicyCLOUD storage: Users, Incentives, Task…
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	all
Role	ROL-02
Source	PolicyCLOUD data repository
Success Criteria	N/A
Expected delivery date	M12 (1st version of prototypes)

TABLE 80: SYSTEM REQUIREMENT REQ- SY-DGMIM-03

Section	Description
ID	REQ- SY-DGMIM-04
Title	Incentive Management component(s) Reusability
Level of detail	System

Type	USE (Usability Requirements), SUP (Maintainability and Support Requirements).
Description	Existing open source solution will be evaluated before developing new Incentives Management component(s).
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	all
Role	ROL-02, ROL-04
Source	
Success Criteria	N/A
Expected delivery date	M12 (1st version of prototypes)

TABLE 81: SYSTEM REQUIREMENT REQ- SY-DGMIM-04

Section	Description
ID	REQ- SY-DGMIM-05
Title	Self-hosting incentive management tool
Level of detail	System
Type	ENV (Operational/Environment Requirements), SUP (Maintainability and Support Requirements).
Description	In case the incentive management features will be accomplished by an existing open source solution, the tool might be self-hosted by the PolicyCLOUD infrastructure.
Additional Information	
Priority	DES (desirable requirement)
Reference Use Case	all
Role	ROL-02, ROL-04
Source	
Success Criteria	N/A
Expected delivery date	M24 (2nd version of prototypes)

TABLE 82: SYSTEM REQUIREMENT REQ- SY-DGMIM-05

5.3 Reusable Models and Analytical Tools

Section	Description
ID	REQ- SY-RMAT-01
Title	Minimum hardware requirements
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	The analytical components require minimum 8 CPU cores and 16G of memory to operate
Additional Information	At least for sentiment analysis, opinion mining and situational knowledge analysis.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	The analytical components should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 83: SYSTEM REQUIREMENT REQ -SY-RMAT-01

Section	Description
ID	REQ- SY-RMAT-02
Title	Define a schema to be used to feed the component
Level of detail	System
Type	DATA (data)
Description	A data schema will be defined to be able to process the data coming from the PolicyCLOUD datastore, or from real-time sources, by the component.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02
Source	N/A
Success Criteria	Data can be correctly processed.
Expected delivery date	M12 (1st version of prototypes)

TABLE 84: SYSTEM REQUIREMENT REQ -SY-RMAT-02

Section	Description
ID	REQ- SY-RMAT-03
Title	Define an interface for the component to be used by PolicyCLOUD (PDT)
Level of detail	System
Type	FUNC (function)
Description	An interface will be developed by defining a set of parameters that should be included when this component is executed by the PDT.
Additional Information	Those parameters are related to the input parameters that the component requires, the type of data source that is being analysed, type of output required to be visualized, and more to be discussed if needed.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-03, ROL-04
Source	N/A
Success Criteria	Analytical component can be executed correctly with different combinations of parameters.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 85: SYSTEM REQUIEREMENT REQ-SY-RMAT-03

Section	Description
ID	REQ- SY-RMAT-04
Title	Create a docker image of the component
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	PolicyCLOUD would use Kubernetes for deploying the components, and because of that, analytical components will be developed as a docker image to be able to be deployed in Kubernetes.
Additional Information	Analytical component could be executed by installing the software directly into a virtual machine, just in case the Kubernetes platform is not available.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02
Source	N/A
Success Criteria	A successful deployment using Kubernetes clustering
Expected delivery date	M24 (2nd version of prototypes)

TABLE 86: SYSTEM REQUIEREMENT REQ-SY-RMAT-04

Section	Description
ID	REQ- SY-RMAT-05
Title	Stream data analysis
Level of detail	System
Type	DATA (data)
Description	The policy maker shall be able to perform analysis (at least opinion mining, sentiment and situational analysis) over continuous data coming from streaming data sources (i.e: social networks)
Additional Information	Their feasibility will depend on the readiness of the acquisition (cloud gateways) and pre-processing (data cleaning, data fusion and data interoperability) components.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	The analytical components: opinion mining, sentiment and situational analysis components should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 87: SYSTEM REQUIEREMENT REQ-SY-RMAT-05

Section	Description
ID	REQ- SY-RMAT-06
Title	Kafka streams messaging
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	Continues data coming from streaming data sources such as Twitter Streaming API shall be published on a Kafka cluster
Additional Information	At least for sentiment analysis, opinion mining and situational knowledge analysis.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02
Source	N/A
Success Criteria	The analytical components should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 88: SYSTEM REQUIEREMENT REQ-SY-RMAT-06

Section	Description
ID	REQ- SY-RMAT-07
Title	Batch data analysis
Level of detail	System
Type	DATA (data)
Description	The policy maker shall be able to perform analysis (at least opinion mining and sentiment analysis) over collection of data persisted in the PolicyCLOUD storage.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	The analytical components: opinion mining, sentiment and situational analysis components should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 89: SYSTEM REQUIEREMENT REQ-SY-RMAT-07

Section	Description
ID	REQ- SY-RMAT-08
Title	Standard API for data base accessing
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	All datastores managed in PolicyCLOUD must provide a standard and common API for the data access/manipulation by the analytical components.
Additional Information	At least for sentiment analysis, opinion mining and situational knowledge analysis.
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	The analytical components should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 90: SYSTEM REQUIEREMENT REQ-SY-RMAT-08

Section	Description
ID	REQ- SY-RMAT-09
Title	External data base analysis
Level of detail	System
Type	DATA (data)
Description	The policy maker shall be able to perform analysis over external data bases.
Additional Information	N/A
Priority	ENH (possible future enhancement)
Reference Use Case	UC#1, UC#2
Role	ROL-01, ROL-02, ROL-04
Source	N/A
Success Criteria	The analytical components: opinion mining, sentiment and situational analysis components should be running with no problems
Expected delivery date	M36 (3rd version of prototypes)

TABLE 91: SYSTEM REQUIEREMENT REQ-SY-RMAT-09

5.4 Policy Development Toolkit

Section	Description
ID	REQ- SY-PDT-01
Title	Analytical tools should expose a REST interface to allow their invocation from the PDT
Level of detail	System
Type	FUNC (function)
Description	The PDT must be able to invoke an analytical tool via a standard REST interface that is being registered into its catalogue
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-02
Source	N/A
Success Criteria	All tools must be accessible and be able to be invoked by the PDT
Expected delivery date	M12 (1st version of prototypes)

TABLE 92: SYSTEM REQUIREMENT REQ - SY-PDT-01

Section	Description
ID	REQ- SY-PDT-02
Title	Analytical tools must register the parameters that are required as input
Level of detail	System
Type	FUNC (function)
Description	Each tool accepts a different type of parameters. These should be retrievable by the PDT in order to guide the user to fill those parameters and invoke the tool accordingly
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-02
Source	Parameters must be serializable and follow a common format
Success Criteria	Analytical tool can be invoked with parameters of the same format
Expected delivery date	M12 (1st version of prototypes)

TABLE 93: SYSTEM REQUIREMENT REQ - SY-PDT-02

Section	Description
ID	REQ- SY-PDT-03
Title	Visualizations of Analytics Results via PDT
Level of detail	System
Type	FUNC (function), L&F (look & feel)
Description	Policymakers should be able to view the results of Analytics via PDT
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-03
Source	
Success Criteria	Policymakers can see the visualization of results from requested Analytics
Expected delivery date	M12 (pilot viz) / M24 (multiple viz) / M36 (Full)

TABLE 94: SYSTEM REQUIREMENT REQ- SY-PDT-03

Section	Description
ID	REQ- SY-PDT-04
Title	Policy Model Editing
Level of detail	System
Type	FUNC (function), DATA (data), L&F (look & feel), USE (usability)
Description	The Policymaker – via PDT - should be able to view the structure of an existing policy model,

Section	Description
	modify and save the policy model (if she/he has the proper rights / ownership). Otherwise view-only functions.
Additional Information	Creation of new Policy Models following system templates, or by copying existing ones
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-03
Source	
Success Criteria	Policy Models Editing capability depending on user rights
Expected delivery date	M12 (view structure) / M24 (Editing) / M36 (full version)

TABLE 95: SYSTEM REQUIREMENT REQ- SY-PDT-04

Section	Description
ID	REQ- SY-PDT-05
Title	User Authentication & Authorization
Level of detail	System
Type	FUNC (function), USE (usability)
Description	PDT User should be able to authenticate using her/his credentials into the system (Login). The content will vary depending on the credentials
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-03
Source	
Success Criteria	A User supplies the credentials and enters (logins) into the platform. Can edit her/his policy models, but e.g. only view other / system policy models.
Expected delivery date	M12 (Authentication) / M24 (Authorization)

TABLE 96: SYSTEM REQUIREMENT REQ- SY-PDT-05

Section	Description
ID	REQ- SY-PDT-06
Title	User Notifications on Analytics Progress
Level of detail	System
Type	FUNC (function), L&F (look & feel), USE (usability)
Description	PDT User should be informed about the status changes in the processing of Analytics Requests.
Additional Information	The user should be informed if an analytics request has failed and the related reason. Also, a time estimation for the completion should be given. Finally, should be notified for the completion of the process.

Section	Description
Priority	DES
Reference Use Case	ALL
Role	ROL-03
Source	
Success Criteria	A policymaker who submitted an analysis request, is being notified as the request is being executed regarding its status and time expectations.
Expected delivery date	M24 (initial) / M36 (Full)

TABLE 97: SYSTEM REQUIREMENT REQ- SY-PDT-06

Section	Description
ID	REQ- SY-PDT-07
Title	User Help
Level of detail	System
Type	L&F (look & feeil), USE (usability)
Description	The PDT should support the policymaker with hints, action descriptions and guides as she/he performs policy creation/modification/verification actions.
Additional Information	The users should also be supported during the selection of policies/ KPIs/ Analytics actions.
Priority	MAN (mandatory requirement)
Reference Use Case	ALL
Role	ROL-03
Source	
Success Criteria	A policymaker can, with relative ease, to explore PDT and performs policymaking tasks
Expected delivery date	M12 (hints) / M24 (descriptions) / M36 (guides)

TABLE 98: SYSTEM REQUIREMENT REQ- SY-PDT-07

5.5 Data Marketplace

The task regarding the data marketplace is starting at a later phase of the project, in Y2, therefore no effort has been spent yet. The requirements for this component have been planned to be included in the second version of this deliverable.

5.6 Ethical Framework

As it has been already mentioned in this section, the Ethical Framework of the platform has not yet been defined, and therefore, there have been no requirements at this phase. The analysis of this framework requires the definition of the platform and the scope of the supported use cases, therefore, it has been planned that this work will be delivered at the second version of the deliverable.

6 Software Requirements

This section provides a list of the initial software requirements for the PolicyCLOUD project. These requirements are related with specific software portions, which can be either a program, a software component, an existing product that will be used as part of the overall platform, or a set of combinations of all the above, that implements a specific functionality and provides a set of capabilities via well-defined interfaces. They may include functional or non-functional requirements imposed by a specific software component that are related with:

- Interfaces exposed by the specific software component that describes the way of interaction with the other software portions
- Performance requirements upon this software portion
- The features that required to be implemented by other components
- Conditions or constraints that the software component should or must take into consideration

The following subsections contain all these software requirements per technological component that will provide an autonomous functionality will consist of a specific software building block in the overall PolicyCLOUD architecture. It is worth to mention that the work that is being carried out by the T2.2, which focus on the design of the overall architecture, takes place in parallel with the work of the elicitation of the user requirements and as a result, it is not clear at this phase the exact types of software components that the platform will be consisted of. Due to this, we list an indicative set of software portions that at this phase, seem to be part of the overall architecture. As the project will progress, an updated version of this deliverable will be released on M12, which will contain an updated set with all the components that are part of the overall architecture, and the list of the software requirements will be refined accordingly. The final list of these requirements is expected to be delivered after the final iteration, and they will be documented in the deliverable that is planned to be released on M22.

6.1 Cloud Provisioning

Section	Description
ID	REQ- SO-CP-01
Title	Provisioning of cloud-based resources to set-up the PolicyCLOUD infrastructure
Level of detail	Software
Type	ENV (Operational/Environment Requirements)
Description	The computing resources of the provisioned cloud infrastructure should be scalable to address the requirements of the selected use cases
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-01, ROL-02
Source	N/A
Success Criteria	Partners will access the PolicyCLOUD infrastructure, operate a distributed K8s cluster as a service for the project, and deploy services and pilot use cases.
Expected	The PolicyCLOUD PaaS and IaaS, will be available for trial at M06. Initial resources

Section	Description
delivery date	allocation will be available from M07-M30. Additional resources will be included to scale-up the set-up from M31 to M36.

TABLE 99: SOFTWARE REQUIREMENT REQ - SO-CP-01

6.2 Cloud Register

Section	Description
ID	REQ- SO-CR-01
Title	Access the PolicyCLOUD IaaS and PaaS
Level of detail	Software
Type	ENV (Operational/Environment Requirements)
Description	Access to the IaaS/PaaS will be available either via GUI or CLI
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-01, ROL-02
Source	• The EGI Federated Cloud infrastructure (IaaS): https://egi-federated-cloud.readthedocs.io/en/latest/federation.html • INDIGO-DataCloud PaaS Orchestrator: https://www.indigo-datacloud.eu/paas-orchestrator https://github.com/indigo-dc/orchestrator
Success Criteria	Partners will access the PolicyCLOUD PaaS and IaaS with federated credentials
Expected delivery date	Access to the PolicyCLOUD PaaS and IaaS will be available from M06.

TABLE 100: SOFTWARE REQUIREMENT REQ - SO-CR-01

6.3 Cloud Gateways

Section	Description
ID	REQ- SO-CG-01
Title	Connection to APIs
Level of detail	Software
Type	FUN (Function)
Description	The PolicyCLOUD Gateway Component should facilitate the connection to appropriately specified APIs, for the retrieval of the information, integrating the corresponding security measures and safeguarding information integrity.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04

Section	Description
Source	-
Success Criteria	Successful connection established to the defined data source.
Expected delivery date	M12 (1st version of prototypes)

TABLE 101: SOFTWARE REQUIREMENT REQ -SO-CG-01

Section	Description
ID	REQ- SO-CG-02
Title	File Parsing
Level of detail	Software
Type	FUN (Function)
Description	Parsing of files (e.g. excel or csv files) should be facilitated for the retrieval of the information, for integrating the corresponding security measures and for safeguarding information integrity.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Successful connection established to the defined data source.
Expected delivery date	M12 (1st version of prototypes)

TABLE 102: SOFTWARE REQUIREMENT REQ -SO-CG-02

Section	Description
ID	REQ- SO-CG-03
Title	Connection to (SQL or No-SQL) Databases
Level of detail	Software
Type	FUN (Function)
Description	The connection to an appropriately specified (SQL or No-SQL) Database should be accomplished in order to achieve the retrieval of the information, the integration of the corresponding security measures and the safeguarding of information integrity.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Successful connection established to the defined data source.
Expected	M12 (1st version of prototypes)

Section	Description
delivery date	

TABLE 103: SOFTWARE REQUIREMENT REQ -SO-CG-03

Section	Description
ID	REQ- SO-CG-04
Title	Configuration
Level of detail	Software
Type	FUN (Function)
Description	The PolicyCLOUD Gateway Component should provide access to a configuration service, facilitating configuration of the connection parameters per connection type and source.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Successful configuration of the connection parameters.
Expected delivery date	M12 (1st version of prototypes)

TABLE 104: SOFTWARE REQUIREMENT REQ -SO-CG-04

Section	Description
ID	REQ- SO-CG-05
Title	Pull Connection Type Support
Level of detail	Software
Type	FUN (Function)
Description	The PolicyCLOUD Gateway Component should support pulling data from external data sources (e.g. through REST APIs) per predefined time intervals.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Successful retrieval of information from the defined data sources.
Expected delivery date	M12 (1st version of prototypes)

TABLE 105: SOFTWARE REQUIREMENT REQ -SO-CG-05

Section	Description
ID	REQ- SO-CG-06
Title	Push Connection Type Support
Level of detail	Software
Type	FUN (Function)
Description	The PolicyCLOUD Gateway Component should support data from external data sources being pushed to the platform per predefined time intervals.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Successful collection and internal push of information and data from the defined data sources.
Expected delivery date	M12 (1st version of prototypes)

TABLE 106: SOFTWARE REQUIREMENT REQ -SO-CG-06

Section	Description
ID	REQ- SO-CG-07
Title	Standardized Interface to other internal PolicyCLOUD components
Level of detail	Software
Type	FUN (Function)
Description	The PolicyCLOUD Gateway Component should facilitate the standardised connection to other internal components of the PolicyCLOUD platform, such as the Data Cleaning Component, the Data Fusion Component, etc. The standardisation of the messages should follow a well-defined and structured format, such as XML or JSON.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-04
Source	-
Success Criteria	Proper specification of message structure.
Expected delivery date	M12 (1st version of prototypes)

TABLE 107: SOFTWARE REQUIREMENT REQ -SO-CG-07

6.4 Incentives Management

Section	Description
ID	REQ- SO-IM-01
Title	Define an interface for the component to set the incentives
Level of detail	Software
Type	FUNC (function)
Description	An interface will be developed to set different fields that instantiate an incentive defined by the policy makers and should be executed by the PDT.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-03
Source	N/A
Success Criteria	An incentive can be defined in PolicyCLOUD.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 108: SOFTWARE REQUIREMENT REQ - SO-IM-01

Section	Description
ID	REQ- SO-IM-02
Title	Being able to manage the incentives defined in PolicyCLOUD.
Level of detail	Software
Type	DATA (data)
Description	The incentives defined should be stored in the PolicyCLOUD datastore in a concrete schema. Moreover, those incentives should be managed and consulted by the PDT.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-03
Source	N/A
Success Criteria	Incentives can be managed in PolicyCLOUD
Expected delivery date	M24 (2nd version of prototypes)

TABLE 109: SOFTWARE REQUIREMENT REQ - SO-IM-02

6.5 Data Cleaning

Section	Description
ID	REQ- SO-DC-01
Title	Standardised Interface to other internal PolicyCLOUD components
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should facilitate the standardised connection to other internal components of the PolicyCLOUD platform, such as the Data Gateway. The standardisation of the messages should follow a well-defined and structured format, such us XML or JSON.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Proper specification of message structure.
Expected delivery date	M12 (1st version of prototypes)

TABLE 110: SOFTWARE REQUIREMENT REQ- SO-DC-01

Section	Description
ID	REQ- SO-DC-02
Title	Error identification
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should facilitate the identification of errors associated with conformance to specific constraints, safeguarding that the data measures compare to defined business rules or constraints.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 111: SOFTWARE REQUIREMENT REQ- SO-DC-02

Section	Description
ID	REQ- SO-DC-03
Title	Conformance to specific data types
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaning process should safeguard the conformance to specific data types (e.g. integer, string etc.).
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 112: SOFTWARE REQUIREMENT REQ- SO-DC-03

Section	Description
ID	REQ- SO-DC-04
Title	Conformance to range constraints
Level of detail	Software
Type	FUNC (Function)
Description	The Data Cleaning process should safeguard the conformance to specific range constraints (min and max values).
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 113: SOFTWARE REQUIREMENT REQ- SO-DC-04

Section	Description
ID	REQ- SO-DC-05
Title	Conformance to predefined values
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaning process should safeguard the conformance to specific predefined values (e.g. values selected from a drop-down list).
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 114: SOFTWARE REQUIREMENT REQ- SO-DC-05

Section	Description
ID	REQ- SO-DC-06
Title	Conformance to regular expression patterns
Level of detail	Software
Type	FUNC (functiom)
Description	The Data Cleaning process should safeguard the conformance to regular expression patterns (data that has a certain pattern in the way it is displayed, such as phone numbers e.g. for text formatting "123-45-6789" or "123456780" or "123 45 6789").
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 115: SOFTWARE REQUIREMENT REQ- SO-DC-06

Section	Description
ID	REQ- SO-DC-07
Title	Conformance to value separation
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard the conformance to separation of values (e.g. complete address in free form field without any indication where street ends and city begins).
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 116: SOFTWARE REQUIREMENT REQ- SO-DC-07

Section	Description
ID	REQ- SO-DC-08
Title	Conformance to cross-field validity
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard the conformance to cross-field validity (e.g. the sum of the parts of data must equal to a whole).
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 117: SOFTWARE REQUIREMENT REQ- SO-DC-08

Section	Description
ID	REQ- SO-DC-09
Title	Conformance to correct value representation
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard the conformance to correct representation of the values.
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 118: SOFTWARE REQUIREMENT REQ- SO-DC-09

Section	Description
ID	REQ- SO-DC-10
Title	Conformance to uniqueness
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard the conformance to uniqueness (data that cannot be repeated and require unique values (e.g. social security numbers)).
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 119: SOFTWARE REQUIREMENT REQ- SO-DC-10

Section	Description
ID	REQ- SO-DC-11
Title	Conformance to mandatory field
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that all the mandatory fields are filled in.
Additional Information	N/A
Priority	MAN (mandatory requirement requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 120: SOFTWARE REQUIREMENT REQ- SO-DC-11

Section	Description
ID	REQ- SO-DC-12
Title	Conformance to specific value length
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that all the filled in values which have specific length constraints, are correctly placed.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 121: SOFTWARE REQUIREMENT REQ- SO-DC-12

Section	Description
ID	REQ- SO-DC-13
Title	Conformance to specific coding standard
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that the appropriate attributes respect their defined coding standard.
Additional Information	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Priority	OPT (optional)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	N/A
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 122: SOFTWARE REQUIREMENT REQ- SO-DC-13

Section	Description
ID	REQ- SO-DC-14
Title	Conformance to value uniformity
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that the appropriate attributes respect their defined value representation.
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 123: SOFTWARE REQUIREMENT REQ- SO-DC-14

Section	Description
ID	REQ- SO-DC-15
Title	Identification of duplications
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should facilitate the identification of duplications that could then be removed facilitating easier and more efficient record management and maintenance.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 124: SOFTWARE REQUIREMENT REQ- SO-DC-15

Section	Description
ID	REQ- SO-DC-16
Title	Automatic field completion
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that the data set provided is fully complete and should empower the automatic filling in of information based on interpolation / extrapolation techniques.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Automatic completion of on-purpose excluded values.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 125: SOFTWARE REQUIREMENT REQ- SO-DC-16

Section	Description
ID	REQ- SO-DC-17
Title	Automatic error correction
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that inconsistencies and errors identified are corrected.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Pandas, NumPy, Scikit-learn, Keras
Success Criteria	Automatic correction of on-purpose included erroneous values.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 126: SOFTWARE REQUIREMENT REQ- SO-DC-17

Section	Description
ID	REQ- SO-DC-18
Title	Data verification
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should safeguard that data provided is accurate, especially referring to erroneous inliers, i.e., data points generated by error but falling within the expected range (erroneous inliers often escape detection).
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python library of Cerberus
Success Criteria	Identification of on-purpose included errors.
Expected delivery date	M12 (1st version of prototypes)

TABLE 127: SOFTWARE REQUIREMENT REQ- SO-DC-18

Section	Description
ID	REQ- SO-DC-19
Title	Data logging
Level of detail	Software
Type	FUNC (function)
Description	The Data Cleaner should keep a log file of all identifications of errors, and especially of all automatic corrections of errors and inclusions of values, to safeguard transparency.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2, UC#3, UC#4
Role	ROL-01, ROL-02
Source	Python libraries of Loguru, Logbook and Python Logging
Success Criteria	Logging of errors identified and of values included.
Expected delivery date	M12 (1st version of prototypes)

TABLE 128: SOFTWARE REQUIREMENT REQ- SO-DC-19

6.6 Data Fusion Linking

Section	Description
ID	REQ- SO-DFL-01
Title	Kubernetes and Spark clusters
Level of detail	System
Type	ENV (Operational/Environment Requirements)
Description	A Kubernetes cluster and a Spark cluster installed on it. The actual hardware requirements for a functional PolicyCloud system depends on the amount of expected data to be ingested and analysed by the system. For the project development the requirement is of 4 VMs each with 16 cores and 128GB memory.
Additional Information	
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2, UC3, UC4
Role	All
Source	Kubernetes, Apache Spark
Success Criteria	Demonstration of use cases with large amount of data, with reasonable performance.
Expected delivery date	Initial version on M12, enhancements on M24, M36

TABLE 129: SOFTWARE REQUIREMENT REQ - SO-DFL-01

Section	Description
ID	REQ- SO-DFL-02
Title	Data streaming framework with initial analytic during data ingest
Level of detail	Software
Type	FUNC (function)
Description	A scalable data streaming middleware framework (e.g. Apache Spark Streaming) with capability to integrate analytic functions to process the ingested data.
Additional Information	The analytic functions will be applied according to registered specification per data source.
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2
Role	All
Source	Apache Spark Streaming
Success Criteria	Demonstration of processing and initial analytic on a registered data source providing value to the policy use case.
Expected delivery date	Initial version on M12, enhancements on M24, M36

TABLE 130: SOFTWARE REQUIREMENT REQ - SO-DFL-02

Section	Description
ID	REQ- SO-DFL-03
Title	Data source & tool registration for streaming analytic
Level of detail	Software
Type	FUNC (function)
Description	A capability to register analytic function and register a data source (with schema / metadata) for streaming analytic by a registered analytic function(s) that support the schema / metadata, and applying the registered analytics during streaming of that data source
Additional Information	The registration will include parameters for the analytic function(s).
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2
Role	ROL-01, ROL-02
Source	Registration analytic tasks for activation in Apache Spark Streaming
Success Criteria	Demonstration of registration of new analytic functions, data source the use it, and policy validation scenario using the function and data source.
Expected delivery date	Initial version on M24, enhancements on M36

TABLE 131 SOFTWARE REQUIREMENT REQ - SO-DFL-03

Section	Description
ID	REQ- SO-DFL-04
Title	Data source & tool registration for regular analytics on data at rest
Level of detail	Software
Type	FUNC (function)
Description	A capability to register analytic function and register a data source (with schema / metadata) that can be a subject to a regular analytic on data at rest (that was already ingested to the system) by a registered analytic function(s) that support the schema / metadata.
Additional Information	The registration will include parameters for the analytic function(s).
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2
Role	ROL-01, ROL-02
Source	
Success Criteria	Demonstration of registration of new analytic functions, data source the use it, and policy validation scenario using the function and data source.
Expected delivery date	Initial version on M24, enhancements on M36

TABLE 132: SOFTWARE REQUIREMENT REQ - SO-DFL-04

Section	Description
ID	REQ- SO-DFL-05
Title	Seamless Analytics on Hybrid Data at Rest
Level of detail	Software
Type	FUNC (function)
Description	Capability of applying analytics seamlessly on data on multiple stores, and mechanism to move older data to long term store.
Additional Information	Specifically, newer (hot) data will be ingested into database while older data will be moved periodically to object storage.
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2, UC3, UC4 (might be demonstrated only in one or more of them).
Role	ROL-02
Source	
Success Criteria	Demonstration of seamless analytic with the data movement process.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 133: SOFTWARE REQUIREMENT REQ - SO-DFL-05

Section	Description
ID	REQ- SO-DFL-06
Title	Data privacy and ownership constraints for multi-tenant analytics
Level of detail	Software
Type	FUNC (function)
Description	The Data Acquisition and Analytics Layer should maintain access control mechanism to respect data privacy stings of the data owner .
Additional Information	Applying analytics to data should be restricted according to the data privacy settings of the data owner.
Priority	MAN (mandatory requirement)
Reference Use Case	UC1, UC2, UC3, UC4 (might be demonstrated only in one or more of them).
Role	ROL-01, ROL-02, ROL-03, ROL-04
Source	
Success Criteria	Validation test of data access restrictions.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 134: SOFTWARE REQUIREMENT REQ - SO-DFL-06

6.7 Data Interoperability

Section	Description
ID	REQ- SO-DI-01
Title	Cleaned Data
Level of detail	Software
Type	DATA (data)
Description	The Data Interoperability component extracts semantic knowledge and good quality information from the cleaned data that will be the input to its system. All cleaned data produced in Data Cleaning Component will be used by Data Interoperability software.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement. Data Interoperability component is mandatory across data lifecycle.
Role	ROL-02, ROL-03, ROL-04
Source	-
Success Criteria	Development of good quality annotated and interoperable data from the provided cleaned data.
Expected delivery date	M12 (1st version of prototypes)

TABLE 135: SOFTWARE REQUIREMENT REQ - SO-DI-01

Section	Description
ID	REQ- SO-DI-02
Title	Triplestore Database
Level of detail	Software
Type	ENV (Operational/Environment Requirements)
Description	Triplestore is needed in order to save correlated, annotated and interoperable data in JSON-LD format and as linked ontologies. Hence, it will be feasible the storage of semantic facts and the support of the corresponding data schema models.
Additional Information	Apache JENA is the preferred Triplestore framework to be used
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement. Annotated and Interoperable data will be saved in the provided triplestore.
Role	ROL-01, ROL-02, ROL-03, ROL-04
Source	Apache Jena Framework
Success Criteria	Successful saving of interoperable data in JSON-LD formats and as linked ontologies.
Expected delivery date	M12 (1st version of prototypes)

TABLE 136: SOFTWARE REQUIREMENT REQ - SO-DI-02

Section	Description
ID	REQ- SO-DI-03
Title	Data Modelling & Ontology Mapping
Level of detail	Software
Type	FUNC (function)
Description	Define the appropriate techniques and tools to map concepts, classes, and semantics defined in different ontologies and datasets and to achieve transformation compatibility through extracted metadata.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement. All cleaned data produced in Data Cleaning Component will be transformed, annotated and mapped by Data Interoperability component.
Role	ROL-02, ROL-03, ROL-04
Source	-
Success Criteria	Successful annotation, transformation and mapping of data and corresponding ontologies in terms of semantic and syntactic interoperability of data.
Expected delivery date	M12 (1st version of prototypes)

TABLE 137: SOFTWARE REQUIREMENT REQ - SO-DI-03

Section	Description
ID	REQ- SO-DI-04
Title	Data Schemas & Data Models
Level of detail	Software
Type	DATA (data)
Description	Define the exact data schemas and models that will be used from the analytical components and will derive and produced by the Data Interoperability Component. Incoming and cleaned data will be modelled and transformed according to the defined schemas and models.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	All 4 Use Cases will be used in order to implement this requirement.
Role	ROL-01, ROL-02, ROL-03, ROL-04
Source	-
Success Criteria	Development of models and schemas corresponding to reference and analytical problems and tasks.
Expected delivery date	M12 (1st version of prototypes)

TABLE 138: SOFTWARE REQUIREMENT REQ - SO-DI-04

6.8 Data Store

Section	Description
ID	REQ- SO-DS-01
Title	Minimum hardware requirements
Level of detail	Software
Type	ENV (Operational/Environment Requirements)
Description	The datastore requires minimum 4G of memory to operate
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	The datastore should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 139: SOFTWARE REQUIREMENT REQ - SO-DS-01

Section	Description
ID	REQ- SO-DS-02
Title	Being able to fragment a dataset and move the data fragments across different nodes.
Level of detail	Software
Type	DATA (data)
Description	The adaptable distributed storage should be able to split a dataset into different regions, and move these regions to different data nodes, in order to adapt in case of increased load (both in terms of user workload or data load) so as to achieve efficient consumption, based on the provided resources.
Additional Information	When a movement (move, split, join) of a data fragment occurs, the storage must not suffer from a down-time. On the contrary, it must remain operational with minimum overhead on the overall performance.
Priority	DES (desirable requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	Data can be moved in different nodes
Expected delivery date	M24 (2nd version of prototypes)

TABLE 140: SOFTWARE REQUIREMENT REQ - SO-DS-02

Section	Description
ID	REQ- SO-DS-03
Title	Provide standard connectivity mechanisms
Level of detail	Software
Type	DATA (data)
Description	The datastore must implement standard connectivity mechanisms to provide access and allow for the query execution
Additional Information	It should provide a JDBC[3] implementation. Additional standard implementation would be the ODAta [4]
Priority	MAN (mandatory requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	All components can retrieve and store data to the data store
Expected delivery date	M24 (2nd version of prototypes)

TABLE 141: SOFTWARE REQUIREMENT REQ - SO-DS-03

[3] https://jcp.org/en/jsr/detail?id=221
[4] https://www.odata.org/

Section	Description
ID	REQ- SO-DS-04
Title	Requirement for a Kubernetes cluster to enable the deployment
Level of detail	Software
Type	ENV (Operational/Environment Requirements)
Description	The infrastructure of PolicyCLOUD should use Kubernetes for deploying the various application/platform components, the adaptable distributed engine must be able to deploy and configure additional data nodes via this technology.
Additional Information	N/A
Priority	DES (desirable requirement)
Reference Use Case	All
Role	ROL-02, ROL-04
Source	N/A
Success Criteria	A successful deployment using Kubernetes clustering
Expected delivery date	M12 (1st version of prototypes)

TABLE 142 SOFTWARE REQUIREMENT REQ - SO-DS-04

6.9 Opinion Mining

Section	Description
ID	REQ- SO-OM-01
Title	Opinion Mining
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe events and social attitude regarding specifics topics (i.e: a policy, a demonstration, a group of people, a wine ..) extracted from datasets and social networks.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 143: SOFTWARE REQUIREMENT REQ - SO-OM-01

Section	Description
ID	REQ- SO-OM-02
Title	Named entity recognition
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to identify specifics entities (users, locations, groups, …) of its interest cited on a text.
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 144: SOFTWARE REQUIREMENT REQ - SO-OM-02

Section	Description
ID	REQ- SO-OM-03
Title	Social media graph analysis
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to identify those users who are talking more about a topic (i.e.: a policy, a demonstration, a group of people, a wine, ...)
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 145: SOFTWARE REQUIREMENT REQ - SO-OM-03

Section	Description
ID	REQ- SO-OM-04
Title	Twitter User Monitoring
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to identify and monitor most popular users (at least on Twitter) who comment about specifics hashtags or topics
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 146: SOFTWARE REQUIREMENT REQ - SO-OM-04

Section	Description
ID	REQ- SO-OM-05
Title	Twitter Hashtags Detection
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to identification of Twitter style hashtags from text
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 147: SOFTWARE REQUIREMENT REQ - SO-OM-05

Section	Description
ID	REQ- SO-OM-06
Title	Twitter Hashtags and Mentions Tacking
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to find and monitor mentions on Twitter regarding specifics hashtags or topics (i.e.: a policy, a demonstration, a group of people, a wine, ...)
Additional Information	N/A
Priority	OPT (optional requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The opinion mining component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 148: SOFTWARE REQUIREMENT REQ - SO-OM-06

6.10 Sentiment Analysis

Section	Description
ID	REQ- SO-SA-01
Title	Social Media Sentiment Analysis
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe the sentiment about what the citizens say in social media channels regarding certain topics.
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The sentiment component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 149: SOFTWARE REQUIREMENT REQ - SO-SA-01

Section	Description
ID	REQ- SO-SA-02
Title	RSS Feed Sentiment Analysis
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe the sentiment in RSS feeds channels regarding certain topics
Additional Information	N/A
Priority	DES (desirable requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	Capturean tool
Success Criteria	The sentiment component should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 150: SOFTWARE REQUIREMENT REQ- SO-SA-02

6.11 Behavioural Analysis

Section	Description
ID	REQ- SO-BA-01
Title	Policy modelling language
Level of detail	Software
Type	FUNC (function)
Description	A special-purpose modelling language needs to be developed that will allow policy practicioners to describe the characteristics of the population on which the policy will be applied and the specific policy mechanisms.
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	At this point it is not clear which use case will be used to showcase the implementation of this requirement
Role	ROL-03, ROL-04
Source	-
Success Criteria	Development of models corresponding to reference problems in network science
Expected delivery date	M12 (1st version of prototypes)

TABLE 151: SOFTWARE REQUIREMENT REQ - SO-BA-01

Section	Description
ID	REQ- SO-BA-02
Title	Behavior simulator
Level of detail	Software
Type	FUNC (function)
Description	Behavioral simulator that accepts as input and runs models developed using REQ-SO-BA-01
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	At this point it is not clear which use case will be used to showcase the implementation of this requirement
Role	ROL-03, ROL-04
Source	-
Success Criteria	Execution of models created in REQ-SO-BA-02
Expected delivery date	M12 (1st version of prototypes)

TABLE 152: SOFTWARE REQUIREMENT REQ - SO-BA-02

Section	Description
ID	REQ- SO-BA-03
Title	User Interface for the Behavioral Analysis component
Level of detail	Software
Type	FUNC (function), L&F (look & feel)
Description	Web-based interface that will allow I/O of population data and policy models along with control of the behavioral analysis component
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	At this point it is not clear which use case will be used to showcase the implementation of this requirement
Role	ROL-03, ROL-04
Source	-
Success Criteria	Web-based interaction with the REQ-SO-BA-02
Expected delivery date	M12 (1st version of prototypes)

TABLE 153: SOFTWARE REQUIREMENT REQ - SO-BA-03

Section	Description
ID	REQ- SO-BA-04
Title	Fault-tolerant and safe operation of the behavioral analysis component in a cloud environment
Level of detail	Software
Type	FUNC (function)
Description	Development of prevention, monitoring, and recovery methods for fault-tolerant and safe operation of the behavioral analysis component
Additional Information	-
Priority	MAN (mandatory requirement)
Reference Use Case	At this point it is not clear which use case will be used to showcase the implementation of this requirement
Role	ROL-02, ROL-03, ROL-04
Source	-
Success Criteria	Stress testing the operation of the behavioral analysis component under extreme load and malicious/unsafe usage scenarios.
Expected delivery date	M24 (2nd version of prototypes)

TABLE 154: SOFTWARE REQUIREMENT REQ - SO-BA-04

6.12 Situational Knowledge Analysis

Section	Description
ID	REQ- SO-SKA-01
Title	Social Media Data Categorization
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe text-based information (coming from social media) classified into defined categories for report generation.
Additional Information	The defined hierarchical classification (categories) will be provided by the use cases.
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	
Success Criteria	The situational knowledge acquisition component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 155: SOFTWARE REQUIREMENT REQ - SO-SKA-01

Section	Description
ID	REQ- SO-SKA-02
Title	RSS Feed data categorization
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe text-based information (coming from RSS Feed) classified into defined categories for report generation.
Additional Information	The defined hierarchical classification (categories) will be provided by the use cases.
Priority	MAN (mandatory requirement)
Reference Use Case	UC#1, UC#2
Role	ROL-02, ROL-04
Source	
Success Criteria	The situational knowledge acquisition component should be running with no problems
Expected delivery date	M24 (2nd version of prototypes)

TABLE 156: SOFTWARE REQUIREMENT REQ - SO-SKA-02

Section	Description
ID	REQ- SO-SKA-03
Title	Supervised predictive analysis
Level of detail	Software
Type	FUNC (function)
Description	The policy maker shall be able to observe predictions based on historic information
Additional Information	As an example, the UC#2: Prediction the quality of the next wine crop
Priority	OPT (optional requirement)
Reference Use Case	UC#2
Role	ROL-01, ROL-02, ROL-04
Source	
Success Criteria	The situational knowledge acquisition component should be running with no problems
Expected delivery date	M12 (1st version of prototypes)

TABLE 157: SOFTWARE REQUIREMENT REQ - SO-SKA-03

6.13 Optimization and Reusability

Section	Description
ID	REQ- SO-OR-01
Title	Use an operational and an analytical database to optimize the query execution
Level of detail	Software
Type	DATA (data)
Description	Data should be ingested to the operational datastore. When they become obsolete and thus, should be considered historical, they would need to be moved to the analytical datastore that can execute queries on BigData more efficiently
Additional Information	N/A
Priority	MAN (mandatory requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Data can be moved from one database to the other and the performance should be improved
Expected delivery date	M24 (2nd version of prototypes)

TABLE 158: SOFTWARE REQUIREMENT REQ - SO-OR-01

Section	Description
ID	REQ- SO-OR-02
Title	Provide access to data stores via a single and common interface.
Level of detail	Software
Type	FUNC (function)
Description	PolicyCLOUD includes two different data stores: the LeanXcale relational data store and IBM object store. The dataset can be fragmented and distributed over the two data stores (historical data being moved to object store). However, the application should be kept unaware of these internal data transfers. The application needs a common interface to submit queries, without having to specify where the data is stored.
Additional Information	A federation mechanism is required that will encapsulate the process of data retrieval from the two data stores. The LeanXcale access point will act as the federator between the relational and the Object Storage. The LeanXcale database already provides a common JDBC interface for data connectivity. The federator will receive the query and execute it in both data stores. For the object store, the access would be via Spark SQL, with the assistance of Apache Hive for storing the metadata of the schema catalogue, which can also be transparently accessible via a JDBC interface. The federator will take into consideration the operations that can be supported in order to push down the operations accordingly. Regarding the relational store, all operations will be pushed down to the store. At the very end, the federator will merge the results and return back the result set. It shouldn't count data that appears in both data stores twice
Priority	MAN (mandatory requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use

Section	Description
Role	ROL-04
Source	N/A
Success Criteria	A query in the common interface can access data that are stored in both stores
Expected delivery date	M24 (2nd version of prototypes)

TABLE 159: SOFTWARE REQUIREMENT REQ - SO-OR-02

Section	Description
ID	REQ- SO-OR-03
Title	Move historical data from the relational data store to the object store.
Level of detail	Software
Type	DATA (data)
Description	Data ingested by the use cases will be stored into the relational datastore, as they are operational, in order to ensure data consistency in terms of ACID properties. After a configurable period, called the freshness window (which depends on the data set), the data becomes outdated and is no longer used by operational workloads. However, this historical data is still valuable and can be exploited by Big Data analytics algorithms. This data should be moved from the LeanXcale data base to the IBM object store.
Additional Information	A mechanism should be implemented that monitors the freshness window and decides whether or not a data movement should take place. The mechanism must allow the data pulling of the data slice from the operational datastore and the persistently storage on the object store. During the data movement, the mechanism should allow the continuous execution of data retrieval from the data federator, so that no down time should be observed, while ensuring the data consistency.
Priority	MAN (mandatory requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Data can be moved automatically from one store to the other
Expected delivery date	M24 (2nd version of prototypes)

TABLE 160: SOFTWARE REQUIREMENT REQ - SO-OR-03

Section	Description
ID	REQ- SO-OR-04
Title	Inform the LeanXcale data store when data are imported to the object store.
Level of detail	Software
Type	FUNC (function)
Description	When data are pulled from the operational datastore, the LeanXcale data base can drop them. However, due to the asynchronous design, the LeanXcale data base cannot know when the data has been made available to the object store. As a result, the object store must inform the LeanXcale data base regarding the successful insertion of the data, so that the LeanXcale data base can safely drop these data
Additional	One possible solution to deal with this requirement will be the introduction of marking the

Section	Description
Information	data to be transferred to the object store by additional timestamps. Data that is being flushed and exported to the object store can be marked that way, so that later, the object store can inform the LeanXcale data base that this bunch of data has been successfully imported. By doing so, the federator component can push down operations accordingly, and only request specific data from the underlying data stores. Data that are known to the LeanXcale database that has been previously uploaded to the object store, will not be retrieved by the federator and can be safely discarded by the vacuum process of the LeanXcale data base.
Priority	MAN (mandatory requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Data moved to object store are now dropped from LeanXcale. A query to LeanXcale will not return any results regarding those data
Expected delivery date	M24 (2nd version of prototypes)

TABLE 161: SOFTWARE REQUIREMENT REQ- SO-OR-04

Section	Description
ID	REQ- SO-OR-05
Title	Optimize query execution
Level of detail	Software
Type	DATA (Data)
Description	The federator receives a query and executes it into the different stores. The federator will be based on the LeanXcale query engine. The latter provides a query optimizer, which allows it to examine the different execution plans that can be produced in order to execute a query. However, it has been implemented to evaluate plans to be executed locally. It should be extended in order to take into consideration the operations that can be pushed down to the object store, and whether or not it is worth for an operator to be pushed down, according to the response time of the execution from Spark SQL, the amount of data that will be retrieved to the federator etc.
Additional Information	As every operation that can be supported by the object store will be pushed down to be executed locally, in order to avoid transferring a big amount of data through the network and process them in the query engine level, the implementation of this requirement corresponds to the following two aspects: the choose of the optimal strategy for executing the JOIN operation concerning data tables that are distributed and split to the two stores, and the redefinition of the query execution plan, in order for the query federator to exploit data locality and reduce the number of rows that will be retrieved and transferred from the object store via the network.
Priority	DES (desirable requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Response time of the query execution is improved

Section	Description
Expected delivery date	M36 (3rd version of prototypes)

TABLE 162: SOFTWARE REQUIREMENT REQ- SO-OR-05

Section	Description
ID	REQ- SO-OR-06
Title	Optimize access to Object Storage.
Level of detail	Software
Type	DATA (Data)
Description	In order to perform analytics efficiently on Object Storage, a client-side caching/acceleration layer is needed. This is critical for a hybrid cloud scenario, where some of the customer data is on premise (potentially the LeanXcale data base and Spark) and some is in the cloud (potentially IBM COS). In such a scenario, when performing analytics, data needs to move from COS to Spark across the WAN, therefore minimizing the amount of data movement when part of the data is retrieved multiple times is of utmost importance. A similar scenario occurs in a multi-cloud environment, where a dataset may be distributed among more than one cloud, also requiring data transfer across the WAN for the purposes of analytics.
Additional Information	This complements data skipping and data layout techniques to further reduce the KPI measuring the number of bytes sent from Object Storage to Spark.
Priority	DES (desirable requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Response time of the query execution is improved
Expected delivery date	M36 (3rd version of prototypes)

TABLE 163: SOFTWARE REQUIREMENT REQ- SO-OR-06

Section	Description
ID	REQ- SO-OR-07
Title	SQL Grammar extension
Level of detail	Software
Type	DATA (Data)
Description	In order to better support the seamless, an extension of the SQL grammar is needed
Additional Information	The grammar extensions will allow the database administrator to define that a data table can be split across the two datastores, and will allow him to provide additional information like the time window of the data slice, along with other configuration attributes like the minimum size of a data slice that is allowed to be moved, time frequency of the moving action etc.
Priority	DES (desirable requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04

Section	Description
Source	N/A
Success Criteria	Data user can use the standard JDBC connection with an extension of the SQL grammar to be able execute DDLs
Expected delivery date	M36 (3rd version of prototypes)

TABLE 164: SOFTWARE REQUIREMENT REQ- SO-OR-07

Section	Description
ID	REQ- SO-OR-08
Title	Ensure data consistency when a moving action is taking place
Level of detail	Software
Type	DATA (Data)
Description	When data is moving from the operational store to the object store, data might either co-exist in both stores, or are non-existed in any store. The framework must be able to serve requests for data retrieval with no downtimes during this process, and the data should be consistent, meaning that the result of the execution of a query should be the same, no matter if the data are being moved.
Additional Information	The operational datastore must not withdraw a data slice, until an acknowledgement of a persistence storage is being notified by the object store. In this case, data can co-exist in both stores. The Query Federator of the framework must take this into account, and re-write the queries to be executed in both stores accordingly in order to scan records on the visible data set in each store. In order to ensure data consistency when parallel transactions are being executed, before, during and after the data moving process, it will rely on the transactional manager of the operational datastore.
Priority	MAN (mandatory requirement)
Reference Use Case	Not clear at this phase which use case will be candidate to showcase it is use
Role	ROL-04
Source	N/A
Success Criteria	Queries return equivalent results as before, during and after the moving of the data slice
Expected delivery date	M24 (2nd version of prototypes)

TABLE 165: SOFTWARE REQUIREMENT REQ- SO-OR-08

7 State of the Art analysis

This section presents the state-of-the-art analysis in the various sectors that the PolicyCLOUD project is being involved. Whenever is possible, it links the state-of-the-art technologies that are described in the following subsections with the context of the project and state how the platform can benefit from the use of those technologies.

7.1 Evidence based policy making and data analytics

Evidence Based Policy Making (EBPM) is a term usually applied when policy choices are performed based on objective evidences using a scientific approach, rather than based on intuition, random, ideology capricious, hidden interests or just theory. Even this approach is known since some centuries ago, it was the Blair administration in UK that brought it back to the political agenda in the late 90s to end "ideologically-based decision making and 'questioning inherited ways of doing things' [41]. EBPM can be used through all the policy making cycle [42], as seen in Figure 3; **1. Agenda setting**, to take decisions on which public issue requires the most attention to take action; in **2. Policy formulation**, to define the strategies that can address the issue in the best way; in **3. Adoption**, to approve the regulatory measures based on objective advice; in **4. Implementation**, to implement the necessary infrastructure following a methodological approach that best supports the policy application; and **5. Monitoring and evaluation**, to assess if policies have reached their targets and therefore are successful or have to be revised.

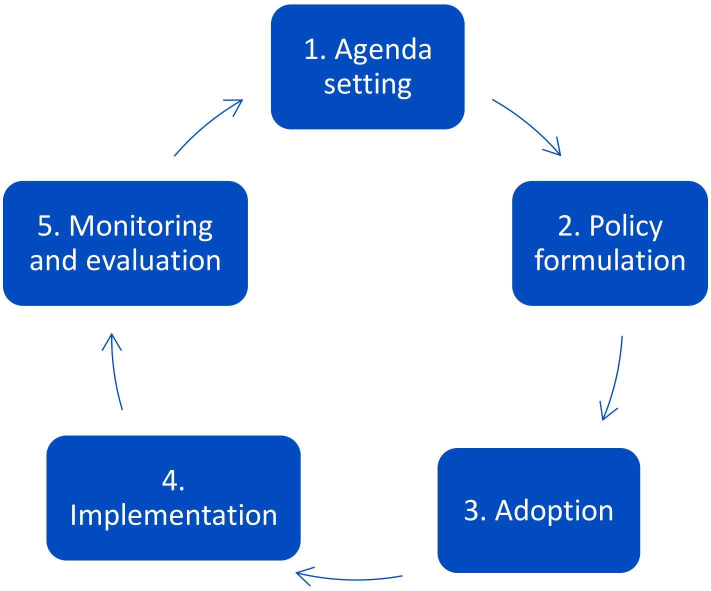


FIGURE 3: POLICY MAKING CYCLE

According to the Roadmap for future research directions from the Big Policy Canvas project [43], more advanced versions of this cycle exist, taking into account the use of big data analytics that enhance the previous cycle by providing evaluation capabilities to the whole cycle, and not only to the latest phase of monitoring and evaluation.

Many challenges are still not solved, as the use of data analytics in the frame of the policy making cycle is not a simple issue. It raises several challenges in the data gathering, integration and reuse. As well, the participation of several types of stakeholders raises many privacy and security issues. Additionally, there are problems in the use of artificial intelligence automation, for example, for biased decision-making as a result of the bias in training data used in this type of systems [43].

In addition to these issues, some bottlenecks and enablers in the application of EBPM have been found [43]:

- Collection of big amounts of data is possible, but quality problems are still a big issue for the use of big data in public policies.
- Resources and budget limitations in public sector are often a burden that must be overcome.
- Interoperability issues with data from several sources, internal and external.
- Leadership issues and the impact of change of political direction after new administrations take over the government.
- Job market availability, limitations of data scientists.
- The importance of having a clear strategy and leadership for the use of data analytics results in the policy making process.
- Providing an opportunity for the update of legacy applications, improving efficiency and interoperability.
- Improving the perception of efficiency of public sector, with high quality services with lower costs.
- Being careful about the use of big data technologies, as they may be a big opportunity for improving the service of public sector, but at the same time could be misused causing negative impacts to the citizens and even erode trust in public authorities.

Six research clusters, or open sets of questions, have been defined for the use of big data technologies in the scope of EBPM, according to Big Policy Canvas project [43], four of them according to the big data value chain and two which are horizontal to the whole value chain, as depicted in Figure 4.

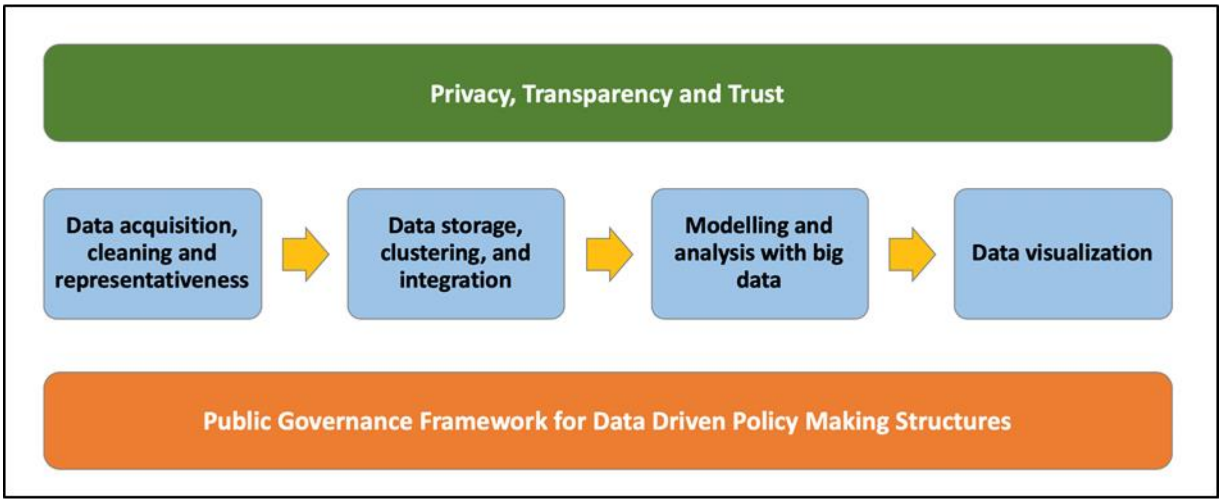

Privacy, Transparency and Trust
Data acquisition, cleaning and representativeness
Data storage, clustering, and integration
Modelling and analysis with big data
Data visualization
Public Governance Framework for Data Driven Policy Making Structures

FIGURE 4: RESEARCH CLUSTERS ACCORDING TO THE BIG POLICY CANVAS PROJECT

The horizontal **Privacy, Transparency and Trust** cluster deals, on one hand, with the issues related to data ownership, security and privacy, and, on the other hand, the issues related to the transparency of the policy making. For the security and privacy, it requires tackling the issues derived from the data ownership and usage, ensuring that data collected is not used for other purposes than the ones for which it is provided. In this regard, research in the area of data regulation and standards for data generated by sensors, devices or in social media is still an open issue. As for the transparency in the policy making, the availability of public open data allows the scrutiny of these policy making processes and the policies implemented and their outcomes. To achieve this, the publication of public data under co-ownership conditions allows the audit and the reuse by society stakeholders. Additionally, algorithms for automated decision making and screening can lead to discrimination, traceability and lastly to breach data protection rules. To avoid this type of misuse, a set of ethical standards have to be developed, ensuring the compliance during their design phase. Moreover, a potential area of research to ensure fairness in algorithms design is open in co-creation approaches for the design of public services when citizens are involved. Questions related to this cluster will be addressed by the Ethical framework and privacy enforcement in the **PolicyCLOUD** architecture as part of **WP3 Cloud Infrastructures Utilisation & Data Governance**, and by the use-cases co-creation approach in **WP6 Use cases Adaptation, Integration & Experimentation**.

The second horizontal cluster, **Public Governance Framework for Data Driven Policy Making Structures**, refers to the set of rules to manage evidence-based policy-making to apply information technology in a way that it is possible to improve the policy-making process and better understand the underlying societal problems that have to be addressed. Implementation through the efficient use of data, achieving a rational, participative and transparent process is key. These aspects are addressed by some of the tasks in **WP5 Cross-sector Policy Lifecycle Management** of **PolicyCLOUD**, like, Cross-sector Policy Lifecycle Management for the modelling and design of policies, collection, experimentation, adaptation, optimisation and implementation of policies and their compliance monitoring.

The third cluster which is the first vertical one is the **Data Acquisition, Cleaning and Representatives**. This has to deal with the huge amount and variety of data sources, according to the origin (own public sector data, social networks, open data, private data), the types of data (internet of things, sensor data, real time data, geo-location, image, video, sound) and the sources (traffic and transport data, administrative processes, citizens, scientific). Meaningful conclusions are drawn up by extracting the behavioural essence associated to all this huge amount of data through the use of big data technologies. Nevertheless, there are some implicit issues with the use of such massive amounts of data in terms of quality, random and systematic errors. The impact of random errors can be minimised by the increase of the data size, but for systematic errors it is not the case. Anyhow, big data is not the Holy Grail that can provide an answer to all the questions, as it is prone to suffer problems due to bias and data inaccuracy, which can be avoided through cleansing mechanisms. These aspects are addressed on **PolicyCLOUD** in **Task 3.3 Cloud Gateways & APIs for efficient data utilisation**, and in **Task 4.2 Enhanced interoperability & Data Cleansing**.

The fourth cluster is the **Data Storage, Clustering, and Integration**. The use of data from so many and different sources, carries the implicit challenge of dealing with data that has not been produced with the goal of being collected for other specific uses. Therefore, the data collected is of heterogeneous nature, some is structured or semi-structured, and sometimes incomplete. For example, social media data is usually applied to sentiment analysis and opinion tracking processes, while it requires a lot of cleansing, and the resulting information is usually biased because of its nature. This is only one example of many data flows being continuously collected by all sorts of systems. Repurposing all this data in the policy-making domain requires data scientists and domain experts' skills to make the right interpretation of the data. In addition, it has to be considered the reuse of existing public sector data, and the availability of methodologies and infrastructures for the storage and processing of such big data. These issues are addressed in **PolicyCLOUD** by **Task 4.1 Cross-sector Data Fusion Linking**.

The fifth cluster is the **Modelling and Analysis with Big Data**. This deals with the approaches for modelling forecast scenarios based on big data, the data modelling and the simulation modelling, and the novel approaches and research being undertaken in this area. In **PolicyCLOUD** this is addressed by the tasks in **WP4** that deal with data analytics, **Task 4.3 Situational Knowledge Acquisition & Analysis**; **Task 4.4 Opinion Mining & Sentiment Analysis**; **Task 4.5 Social Dynamics & Behavioural Data Analytics**; and in **WP5** with **Task 5.2 Modelling & Design of Policies**.

Finally, the sixth cluster is the **Data Visualization**. The presentation of information from big data is a challenging issue, so the insights extracted from the data are presented in a meaningful way to humans. It is also relevant in the policy making context so to understand the problems from the results obtained from modelling and analytics tools. In this regard, the provision of evidences based on the identification of KPIs and their relations is key. The most relevant approach for **PolicyCLOUD** data visualisation is the provision of dashboard visualisation to measure and monitor relevant indicators with respect to the final objectives for the corresponding policies. Other methods correspond to info-graphic presentations and visual analytics. The visualisation approach is addressed by **Task 5.3 Policy Development Toolkit including Data Visualisation**.

7.2 Policy interoperable datasets

Nowadays, policy makers publish an increasing amount of their data on the Web in an effort with double fold meaning. In one hand, to comply with the emerging Open Data movement and in the other hand in order to optimize and improve their policy management and development lifecycle. A key to realizing the open data and providing advanced open policies is the ability to merge divergent data and datasets. Hence, interoperability is the key "back office" element across the whole policy making lifecycle and open data semantics [8]. Achieving true interoperability entails different representations, purposes, and syntaxes and will enable improved access to records, datasets and policies. Recent years many approaches, standards, ontologies and vocabularies have been proposed as means of achieving various tasks of interoperability between heterogenous and independent datasets. One of the first approaches on dataset interoperability is the Information Modelling and Interoperability (IMI) model, which further splits the interoperability into three distinct layers: the syntax layer, the object layer and the semantic layer [9]. Likewise, more recently the European Commission, through their program ISA^2 has defined the European Interoperability Framework (EIF) which defines interoperability across the above four layers: (i) organizational interoperability, (ii) semantic interoperability, (iii) technical interoperability and (iv) legal interoperability [10]. In addition, within LOD2 project the NIF framework was designed, which is based on a Linked Data enabled URI scheme for identifying elements in (hyper-)texts and an ontology for describing common semantic terms and concepts of NLP tools and services [11]. An emerging research direction entails automatically discovering links between datasets using Word Embeddings and other components that find links based on syntactic and semantic similarities [12]. Moreover, a recent research focused on implementing a vocabulary (i.e. VoIDext) to formally describe virtual links in order to enable interoperability among different datasets [13]. By defining virtual links with VoIDext RDF schema and by providing a set of SPARQL query templates to retrieve them, the research team achieved to facilitate the writing of federated queries and knowledge discovery among federated datasets. Furthermore, another project in the archaeological domain, also, highlights the use of RDF schemas to achieve dataset interoperability by extracting and exposing archaeological datasets (and thesauri) in a common RDF framework assisted by a semi-automatic custom mapping tool [14]. In addition, a relevant research introduced three metrics to express the interoperability between two datasets: the identifier interoperability, the relevance and the number of conflicts [15]. Another commonly used technology for achieving and enhancing interoperability is the JSON for Linking Data (JSON-LD) format, that has been a W3C recommendation since 2014 to promote interoperability among JSON-based web services [16]. A research in the biological sector highlights the usage of a JSON-LD system, which provides a standard way to add semantic context to the existing JSON data structure, for the purpose of enhancing the interoperability between APIs and data [17]. PolicyCLOUD project will enhance interoperability based on data driven-design, coupled with linked data technologies (e.g. JSON-LD and RDF) and standards-based ontologies and vocabularies to improve both semantic and syntactic interoperability. Moreover, a data modelling by standard metadata schemas will be defined in order to specify the metadata elements that should accompany a dataset within a domain. To this end, linked data will work as the foundation of a common export format for data within PolicyCLOUD Marketplace.

7.3 Enhanced visualizations providing actionable insights

Visualizations are the most understandable way for humans to show the results of data analysis. Well known is the saying: "An image is worth a thousand words", and following this idea, through different kinds of visualizations, images allow heterogeneous users to obtain in a concise, ordered and structured way, a broader knowledge of the information they need in each time, and, consequently, a better decision making. In this line, charts, for example, are a kind of visualization that is widely used because of its easy representation and users' understanding.

In order to visualize data as a chart in a web site there are three different approaches:

- **Use products/tools/software created for that use:**

In the case of products, there are some software tools in the market that have been designed to visualize data according to the most common needs of companies. Some of these products are:

- Datapine[5]: A Software as a Service (SaaS) platform that can be used to display data as charts.
- Microsoft Excel : Microsoft excel is a spreadsheet software developed by Microsoft where there is the possibility to create several different types of charts.
- Grafana[6]: An open source analytics solution that allows us to visualize data in order to understand the trends of it.
- Tableau[7]: Offers a platform to display charts.
- Microsoft PowerBI[8]: Part of the Microsoft Office 365 package; it offers several charts to display the information of a company.
- QlikView[9]: An End-to-End Data integration and analytics tool.

- **Create charts into the backend of the site as image and display them into the front-end:**

Backend libraries are used by the backend of web sites to create charts when it receives the petition. Some of these libraries are:

- JpGraph[10]: A PHP library, valid for PHP5 and PHP7 that can create several types of charts.
- Matplotlib[11]: A Python 3 library that can be used to create static, animated, and interactive visualizations.

- **Create dynamic charts in the front-end on the fly using JavaScript libraries:**

JavaScript Libraries to be used in a front-end.

In the scope of the project, the chosen way to do it is to create a variety of dynamic charts in the front-end, on-the-fly, when end users wish to display the data. This way is lighter for the server and the results are more

[5] https://www.datapine.com/
[6] https://grafana.com/
[7] https://www.tableau.com/
[8] https://powerbi.microsoft.com/
[9] https://www.qlik.com/
[10] https://jpgraph.net/
[11] https://matplotlib.org/

attractive, useful and understandable for users, mostly because of its interactive feature. The chosen language is JavaScript (JS), the most used language for webs. Following this approach there are several JavaScript libraries that can be used to create charts. These libraries use one of these rendering technologies: HTML5 Canvas, SVG (Scalable Vector Graphics) or VML (Vector Markup Language) to create charts. This technique usually needs to request data from the backend by APIs that return data in JSON formats.

Some of the most popular chat libraries are shown in the table below:

Library name	Main Site	License	Rendering technology	Public code repository
amCharts	https://www.amcharts.com/	Proprietary	SVG and VML	https://github.com/amcharts/amcharts4
AnyChart	https://www.anychart.com/	Proprietary	SVG and VML	https://github.com/AnyChart/AnyChart
C3.js	https://c3js.org/	MIT	SVG	https://github.com/c3js/c3
Chartist.js	https://gionkunz.github.io/chartist-js/	WTFPL or MIT	SVG	https://github.com/gionkunz/chartist-js
Chart.js	https://www.chartjs.org/	MIT	Canvas	https://github.com/chartjs/Chart.js
D3.js	https://d3js.org/	BSD-3	SVG	https://github.com/d3/d3
Flot	https://www.flotcharts.org/	MIT	Canvas	https://github.com/flot/flot
FusionCharts	https://www.fusioncharts.com/	Proprietary	SVG and VML	https://github.com/fusioncharts/
Google Charts tools	https://developers.google.com/chart	Free	Canvas, SVG and VML	
Highcharts	https://www.highcharts.com/	Proprietary	SVG and VML	https://github.com/highcharts
Plotly.js	https://plot.ly/javascript/	MIT	SVG	https://github.com/plotly/plotly.js
Ngx-charts	https://swimlane.github.io/ngx-charts/#/ngx-charts/bar-vertical	MIT	SVG	https://github.com/swimlane/ngx-charts

TABLE 166: MOST POPULAR JS CHART LIBRARIES [19], [20], [21], [22], [23]

When developing visualizations, using a framework usually helps to save time and efforts, as they can facilitate the whole process. JavaScript frameworks make it easier developing with JavaScript. Some of the most popular JavaScript frameworks are shown in the table below:

Framework	Main Site	License	Current version	Size
React	https://angular.io/	MIT	v9.1.0	143K
Vue	https://reactjs.org/	MIT	v16.13.1	43K
Backbone	https://vuejs.org/	MIT	v2.6.11	23K
Ember	https://backbonejs.org/	MIT	v1.4.0	7.3K

Framework	Main Site	License	Current version	Size
Meteor	https://emberjs.com/	MIT	v3.17	95K
Nodejs	https://www.meteor.com/	MIT	v12.16.1	
Mithril	https://nodejs.org/	MIT	v13	
Polymer	https://mithril.js.org/	MIT	v2.2.0	

TABLE 167: MOST POPULAR JS FRAMEWORKS [24], [25]

In the table below it is shown which libraries from above table, can be integrated easily in with each framework:

	Angular	React	Vue	Back bone	Ember	Meteor	Nodejs	Mithril	Polymer	Total
amCharts	X	X	X		X	X	X		X	**7**
AnyChart	X	X	X		X	X	X			**6**
C3.js	X	X	X	X	X	X	X	X	X	**9**
Chartist.js	X	X	X			X	X		X	**6**
Chart.js	X	X	X	X	X	X	X	X	X	**9**
D3.js	X	X	X	X	X	X	X	X	X	**9**
Flot	X	X	X		X	X	X		X	**7**
FusionCharts	X	X	X	X	X		X		X	**7**
Google Charts tools	X	X	X	X	X	X	X		X	**8**
Highcharts	X	X	X		X	X	X		X	**7**
Plotly.js	X	X	X		X		X			**5**
ZingChart	X	X	X	X	X		X			**6**
Ngx-charts	X									
Total	**13**	**12**	**12**	**6**	**11**	**9**	**12**	**3**	**9**	

TABLE 168: COMPATIBILITY BETWEEN JS LIBRARIES AND FRAMEWORKS

As can be seen in the previous table, four frameworks have more JavaScript chart libraries compatibilities: AngularJS, React, Vue and NodeJS.The final decision about which framework and which JavaScript library will be used will not take only this information into account, but also if there are any other project requirement that affect directly this decision.

7.4 Re-usability of analytical tools/models decoupled from infrastructure

It is clear today that the modern cloud native architecture should be based on the micro-services pattern[12] where the application is decomposed into modular independent components which consume each other's API to provide the overall functionality. The definite major winning technology today for achieving this decomposition is containers. In addition to the decomposition of application components, containers decouple the application from the underlying infrastructure. Packaging within containers ('Containerization') provides dramatic simplification and speed of deployment on any cloud infrastructure, as well as avoidance of lock-in to any platform. Recoverability, elasticity and scalability properties of the underlying platform are additional benefits of the decoupled architecture. For containers, the dominate core container technology is Docker[13] and the container framework technology is Kubernetes[14].

For analytical tools, specifically over big data, the advances in data center networking make the disaggregation of storage and compute the widely employed pattern today. The decoupling of the storage infrastructure provides even greater advantage over the general application decomposition due the complexity and price of storage platforms, having the ability to re-use and apply any analytical tool on big data residing on any storage platform. An excellent example is Apache Spark[15] , today the most popular open source for big data analytics that supports various analytic frameworks as GraphX for graph processing, MLlib for machine learning, SQL and streaming. Each of its analytic frameworks can work with any storage platform (as HDFS, S3 Object Storage) decoupled from the computing resources. This enables to decouple the analytic logic and modelling from the cumbersome setting and integration details for each platform.

Emerging specialized data format frameworks provide today intermediate layer for the analytics' logic and modelling, exploiting seamlessly features of underlying storage platforms. The major ones today are Delta Lake[16], Apache Iceberg[17] and Apache Hudi. These frameworks provide abstraction over table/file formats for analytics tools, with consistency and performance optimization features as data catalog / meta data, schema and layout evolution, time travel, atomicity, Merge on Read, Copy on Write, limited transactional operations and exploitation of columnar formats as Parquet[18]. A layer above the data format frameworks is the data warehouse / data lake frameworks which provide consistent and controlled access to various data sets and data sources, as Snowflake[19] which hide the actual cloud infrastructure (runs on Amazon, MS Azure, Google Cloud Platform) and Dremio[20] [21]which provides optimized data lake engine based on Apache Arrow for in-memory columnar data processing. These abstraction layers enable data scientists to concentrate on the analytic algorithms and models, and reuse them over different compute and storage infrastructures in much greater ease than in the past,

[12] https://www.ibm.com/cloud/learn/microservices
[13] https://www.docker.com
[14] https://kubernetes.io
[15] https://spark.apache.org
[16] https://delta.io
[17] https://iceberg.apache.org
[18] https://parquet.apache.org
[19] http://pages.cs.wisc.edu/~remzi/Classes/739/Spring2004/Papers/p215-dageville-snowflake.pdf
[20] https://www.dremio.com
[21] https://github.com/dremio/dremio-oss

although of course the picture is not perfect and migration from one platform to another is usually not completely transparent.

7.5 Polyglot analytical tools federating heterogeneous sources and stores

Accessing heterogeneous data sources (a concept often addressed by data integration systems or multidatabases [36], [39]) is a problem that has been widely studied in the literature and with the recent emerge of cloud databases and big data processing, it has been evolved towards polystore systems. The latters provide a common accessibility method in order to retrieve data from a variety of heterogeneous target data stores, such as typical relational DBMS, NoSQL or NewSQL datastores, or HDFS datalakes, involving data that can be either structure, semi-structure or fully unstructured. Their early implementations [34], [35], [38] relied on a single common model that the target datastores had to transform their schema to. A further improved presented by the polystore BigDAWG [31], [32] which defines *islands of information*, where each island is related to a specific data model and language in order to provide access to the underlying data store. It additionally provides the support for queries spanning among the different data models by moving the intermediate datasets between those islands. Moreover, Myria [40] uses a shard-nothing parallel architecture for data federation across the heterogeneous models and query languages and exploits its extended relational model and its unique imperative language for defining transformation rules that will allow the input query to applied to a target datastore-specific call. It is worth to mention that other polystore solutions [29], [30], [33] rely on the application requirements themselves to decide the optimal data placement and the query execution plan.

Spark SQL [28] is a parallel SQL engine that offers tight integration between traditional relational and procedural processing via a standard API, taking advantage of massive parallelism. It offers a DataFrame API that translates relations into arbitrary object collections, thus supporting operations targeting external datastores that transform the data into those collections. It makes uses of data connectors implemented for each supported datastore whose role is to map a data item into this DataFrame. Presto [37] on the other hand is a distributed SQL query engine which makes us of interactive analytic queries against the target datastores. When it comes to the query execution, it allows for massively parallel processing, consisting of a coordinator and multiple workers, each one of those is making use of target specific connectors which implements a common interface. The implementation of the latter encapsulates the target database details on how to access the data source, while provides the data schema metadata to the coordinator to be taken into account during the query plan of the execution. What is more, Apache Drill [26] is another distributed query engine for large-scale datasets that is capable of querying data coming from a various data sources via its own plugins implemented for each one of the latters. Is also uses massively parallel processing that allows for scaling to thousands of nodes while maintaining overall latency, even if when processing petabytes of data. Each of its workers, called *drillbiti* in its terminology, receives a query and compiles it accordingly and decides over an optimized query execution plan that can be parallelized taken into account data locality. Finally, Impala[27], which also provides a massively parallel processing engine, ensures overall low latency and high concurrency for analytical queries, making use of data specific connectors that transforms the retrieved data from an external dataset into Hadoop compliant format, and then makes use of MapReduce jobs combining the intermediate results.

All of the aforementioned solutions that can be considered as polyglot analytical tools that enables the federation of heterogeneous datastores relies on their own specific data model and query language for query execution, and provide technology-specific interfaces for the integration with the applications and the data user analytical tools. PolicyCLOUD will rely on the engine of its central repository provided by LXS, which not only enables for query parallelism with the external datastores, but also supports the combination of massive parallelism with native queries and the optimizability of bind joins, which is addressed by the LeanXcale distributed query engine.

7.6 Efficient data fusion from various data sources

For PolicyCloud scenarios, two aspects of efficient data fusion are important: (1) scalability of massive data ingest from multiple data sources, where burst of incoming metric data may lead to analytic results of required urgent alert in respect to some policy validation rule, and (2) the capability to apply analytics over incoming data in flight, with the intent of storing only the resulted insight rather than the whole bulk data.

For the first aspect, scalable data fusion frameworks are mostly deployed for IoT scenarios as smart cities, where incoming data from multiple sensors or other IoT devices needs to be efficiently processed and stored. These frameworks are classified by several categories ([1]) :

- **Objectives**: Fixing problematic data [3], Improving data reliability, Increasing data completeness
- **Techniques**: Data association – correlation between sources, Increased state estimation by inspecting multiple sources, Prediction, Unsupervised ML, Dimension reduction – for feature extraction e.g. in PCA
- **Data Input and output types**: Data2data, Data2feature, Feature2feature
- **Data source types**: Physical – as temperature or air quality sensors, Cyber – internet sources as web access data and social network data, Participatory – crowdsourcing from data contributed by personal devices [5], Hybrid – data obtained from mixed types of sources [6]
- **Scale**: Sensor level [4], Building wide [7], Inter-buildings, City wide, Inter-city
- **Platform Architectures**: Edge - data sources are processed at the edge, Fog - data sources are processed at a middle layer between the edge and the cloud, e.g. at cloud gateway, Cloud - data sources are processed in the cloud, this is the most common technique practiced by industry and research institutes for processing big data, Hybrid - processing is done in two or more layers (edge, fog and cloud) [2]

For the second aspect, i.e. applying analytics over incoming data in flight with the intent of storing only the resulted insight, there are several tolls, most of them are open source or have open source version. One of the major commonly used tools today that enables analytics on streaming data is Apache Spark Streaming[22] which enables to apply core Spark analytics within live stream processing. It supports various streaming sources as Kafka, Flume, Kinesis, TCP sockets, and the processed data can be pushed to filesystems, databases, or dashboards. Internally Spark Streaming divides the incoming streams into batches, which then can be processed by regular Spark and generate stream of batch results.Another emerging streaming processing tool is KSQL[23] which is open source and Confluent KSQL[24] which is extended commercial version. It provides SQL interface for stream processing above Apache Kafka[25], and even the open source version is designed for mission-critical and scalable deployments. It provides a very simple programing interface (relative to Spark), and supports numerous streaming operations, including data filtering, transformations, aggregations, joins, etc. Other tools are Flume[26] which is supported in many commercial Hadoop distributions, Apache NIFI[27] that is used for data processing

[22] https://spark.apache.org/streaming
[23] https://github.com/confluentinc/ksql
[24] https://www.confluent.io/product/ksql
[25] https://kafka.apache.org
[26] https://flume.apache.org
[27] https://nifi.apache.org

among multiple sources and targets, Apache Storm[28] that is used in many real-time deployments, and Amazon Kinesis[29] for real-time data analytics in the AWS eco-system.

7.7 Efficient cloud infrastructures

The revolution in information technologies we are facing over the past decades has played a decisive and unprecedented role in the development of society, science, technology, and economics. Today we are living in the big data era. Data volume is continuously increasing, doubling every 3 years. Within one minute, 400 hours of videos are uploaded on YouTube, 3.6 million Google searches are conducted worldwide each minute of every day, more than 656 million tweets are shared on Twitter, and more than 6.5 million pictures are shared on Instagram each day. When a dataset becomes so large that its storage and processing become challenging due to the constraints of existing tools and resources, the dataset is referred to as big data. When dealing with huge data volumes to be analyzed, cloud compute and big data come to play as they provide a solution which is both scalable and accommodating for big data analytics. Through hardware virtualization, cloud computing provides the option of storing significant amounts of data with the help of scalability, fault tolerance and availability. This allows Big Data to be available, scalable and fault tolerant through cloud computing. From a technical perspective, cloud computing consists of three **service models**, which can be offered across three **different deployment models**. The service models consist of Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS):

- Software as a Service, also known as cloud application services, represents the most commonly utilized option for businesses in the cloud market. SaaS consists of firms offering the capability to use software applications that are housed off the user's premises. SaaS utilizes the internet to deliver applications, which are managed by a third-party vendor, to its users. A majority of SaaS applications run directly through web browsers, which means they do not require any downloads or installations on the client side.
- Cloud platform services, also known as Platform as a Service (PaaS), allow users to develop their own Web-based applications or to customize existing applications using one or more programming languages and development tools. These services can be considered natural extensions of individual computer platforms. All servers, storage, and networking can be managed by the enterprise or a third-party provider while the developers can maintain management of the applications.
- Cloud infrastructure services, also known as Infrastructure as a Service (IaaS), allow customers to access the equipment and hardware needed to perform computing operations, including storage, processing and networking components.

The three deployment models through which these services can be provided, separately or together, are: (1) private clouds, (2) public or 'community' clouds, or (3) hybrid structures. Private clouds are exclusive to a single user. Public or community clouds are available to the general public or shared by large diverse groups of customers, and hybrid clouds combine public and private elements in the same data center.

For PolicyCLOUD a federated IaaS cloud infrastructure will be set-up for helping policy makers, public authorities and different stakeholders to analyse a wide plethora of datasets collected from different data sources, and provide decision making support to public authorities for policy modelling, implementation and

[28] http://storm.apache.org
[29] https://aws.amazon.com/kinesis

simulation, as well as for policy enforcement and adaptation. Further details about the federated IaaS-type cloud infrastructure used during the project will be provided in D2.2.

7.8 Data Governance Model, Protection and Privacy Enforcement

The data governance model and the tools for protection and privacy enforcement will be used to protect data and ensure decisions across the complete path follow specific guidelines and legislations. Although we are currently not able to precisely define the data governance model and its attributes, we have examined the relevant efforts on models, standards and frameworks for achieving our objectives.

The Data Governance Model will be based on the usage of the Responsibility assignment matrix (RACI) model and will be used to ensure access of different entities to the corresponding datasets at specific phases of the data and the policy lifecycles. RACI is used for clarifying and defining roles and responsibilities in cross-functional or departmental projects or business processes, and it builds on four key responsibilities most typically used. The responsibilities used are the following;

- **Responsible** refers to the people who do the work to complete the task.
- **Accountable** is a single person specified answerable for the correct and thorough completion of the deliverable or task and approves work that responsible provides.
- **Consulted** refers to people whose opinions are taken into account and with whom there is a two-way communication
- **Informed** refers to people that kept up-to-date on progress, often only on completion of the task or deliverable; and with whom there is just one-way communication.

RACI model defines roles and people: a role is a descriptor of an associated set of tasks; may be performed by many people, and one person can perform many roles.

For the PolicyCLOUD Data Governance Model, we have to adapt RACI to be able to model the access of specific stakeholders to specific data at specific points in the lifecycle. For achieving this we have to use an Access Control Mechanism, and in specific Attribute-based access control (ABAC). Access Control Mechanisms are mechanisms realizing various logical access control models that provide the framework and set of boundary conditions upon which the objects, subjects, operations, and rules may be combined to generate and enforce an access control decision.

Several models and mechanisms, with each having its own advantages and limitations. For the sake of completeness, the most dominant ACMs will be listed:

- **Discretionary Access Control (DAC)** where the owner of the object specifies which subjects can access the object. Most operating systems such as all Windows, Linux, and Macintosh and most flavors of Unix are based on DAC models.
- **Mandatory Access Control (MAC)** where the system (and not the users) specifies which subjects can access specific data objects. The MAC model is based on security labels. Subjects are given a security clearance (secret, top-secret, confidential, etc.), and data objects are given a security classification (secret, top-secret, confidential, etc.).
- **Identity Based Access Control (IBAC)** uses mechanisms such as access control lists (ACLs) to capture the identities of those allowed to access an object. In the IBAC model, the authorization decisions are made statically prior to any specific access request and result in the subject being added to the ACL.

- **Role-based access control (RBAC)** employs pre-defined roles that carry a specific set of privileges associated with them and to which subjects are assigned.

- **Attribute-based access control (ABAC)** uses attributes, and policies that express boolean rule sets that can evaluate many different attributes before allowing access. ABAC, therefore, avoids the need for capabilities (operation/object pairs) to be directly assigned to subject requesters or to their roles or groups before the request is made. IBAC and RBAC can be seen as special cases of ABAC, with IBAC using the attribute of "identity" and RBAC using the attribute of "role".

The adaptability and expressiveness of ABAC make it ideal for protecting the data in the lifecycle of PolicyCloud. The key standards that implement ABAC are OASIS standard of extensible Access Control Markup Language (XACML)[30] and the Abbreviated Language For Authorization (ALFA). XACML uses XSD notation in order to model the three basic artefacts (policy, the request and the response) which are required in an authorization scenario. ALFA is a pseudocode language that respects the XACML model (contains the same structural elements as XACML i.e. PolicySet, Policy, and Rule), but uses JSON instead of XML for the definition of access-control policies and maps directly into XACML. More information about ABAC and XACML will be provided in the deliverables D3.1/D3.4/D3.7.

For the PolicyCloud Data Governance Model, we have to identify a set of properties (as part of the ABAC concept) 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).

Finally, regarding the actual implementation of the privacy enforcement mechanism, the ABAC based authorization should be performed through the evaluation of policies per each data access request. Different tools will be examined in order to select, adapt and extend the most appropriate for PolicyCloud, namely PaaSword[31], Drools[32], Keycloak[33] and WSo2 Balana[34], AuthzForce[35], based on the trade-off between flexibility, expressivity support and efficiency.

7.9 Cross-sector Policy Lifecycle Management

While several application domains are exploiting the added-value of analytics over various datasets to obtain actionable insights and drive decision making, the public policy management domain has not yet taken advantage of the full potential of the aforementioned analytics and data models. Diverse and heterogeneous datasets are being generated from various sources, which could be utilized across the complete policies lifecycle (i.e. modelling, creation, evaluation and optimization) to realize efficient policy management. Although it is imperative that policymaking is based on scientific evidence, in many countries, particularly low- and middle-income countries, evidence-informed decision-making remains the exception rather than the rule [44]. Even into high-income countries internal data, reports and the opinions of internal staff members are the kinds of information used most frequently instead of research evidence [45][46].

[30] http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html
[31] https://paasword.io/
[32] https://www.drools.org/
[33] https://www.keycloak.org/
[34] https://github.com/wso2/balana
[35] https://authzforce.ow2.org/

Agent-based dynamic simulation platforms to identify beneficial policies and interventions have been recently reported for cases such as the impact of sugar-sweetened beverage warning labels [47], the relation of urban crime with obesity [48], and reducing alcohol-related harms [49]. The platforms take residential and sociodemographic data and, through experimental scenarios, estimate the probability and the evolution of various factors. The participation of stakeholders in running simulations and different scenarios builds an [50].

Recently, many projects have been spawned in the direction of evidence-based policymaking via the effective use of big data analytics. We can divide them into two categories. Into the first category, heterogenous big-data datasets are collected, even real-time, to produce quantitative evidence supported by what-if scenarios. Such projects are BigO[36] (Big data against childhood obesity), and MIDAS[37] (Meaningful Integration of Data, Analytics and Services).

In the second category there is an additional layer, where the scientific evidence is framed in a way to support the formulation of public policy models and their management. The EVOTION [38] Project (Big data for hearing loss interventions), is one of the few attempts with specific outcomes, to formulate evidence-based policies. PHP decision making (PHPDM) models are structures having the following set of building elements: Goals, Objectives, Decision Criteria, Data, Factors, Types of Analysis and Policy Actions [51]. The ontology instance of the PHPDM is compiled through a reasoner, producing the corresponding Big Data Analytics (BDAs) components for the delivery of quantitative results [39]. CrowdHEALTH[40] is an international research project co-funded by the European Commission that integrates high volumes of health-related heterogeneous data from multiple sources with the aim of supporting policymaking decisions [52]. The front-end of the platform is a health policy creation and evaluation environment, which provides advanced decision support, through data-driven analytic tools, both in aggregate as well as in personalized fashion. It presents a modular architecture and a secure big data processing workflow [53], while the Public Health Policy Model (PHPM) structure has elements consisting of Actors, Stakeholders, Key Performance Indicators KPIs, Formula (for the computation of the KPIs, Data, and Health Analytics Tools [54].

From the reported early attempts to develop platforms assisting policymakers to benchmark, simulate and forecast outcomes of policy decisions, we can discern challenges towards many directions, some of which are listed here:

- Representing a policy with measurable and quantitative variables.
- Finding, collecting, converting and handling big data sources at spanning time scales.
- Covering sensitivity of personal data, security and trustworthiness.
- Distributed reusable Big Data Analytics independent of cloud vendors, architectures or analytics frameworks.
- Full tracking and versioning of developed Policy Models along with supporting evidence and confidence intervals/error metrics.
- Hiding technical complexity, providing easy interaction of the policymaker with the platform.

To address these challenges. along with the complexity of the policy formulation and lifecycle management, there are many ICT tools spanning across categories: Visualization tools, Argumentation tools, eParticipation tools,

[36] https://bigoprogram.eu/
[37] http://www.midasproject.eu/
[38] http://h2020evotion.eu/
[39] https://scite.ai/reports/towards-a-model-driven-platform-for-6Pm3Jr
[40] https://crowdhealth.eu

Opinion mining tools, Incentive Management Tools, Simulation tools, Serious games, Persuasive tools, Social network analysis (SNA) tools, Big data analytics tools, and Semantics & linked data tools [55].

In a recent work, Giabbanelli et al. [56] highlight five areas where policy-making supporting software should assist:

- Participants access and update supporting definitions and evidence.
- Manually exchange of information between the new software and visualization, argumentation, and simulation tools.
- Iterative process to discern policy ‘inputs’ within context (loops in cognitive maps).
- Ability to monitor the outcomes of interventions via disjoint paths.
- Finding and filtering rippling effects of interventions.

Summarizing, the latest developments in big data analytics and the vast amounts of data that are being generated by different sources provide an opportunity for optimizing cross-sector policy lifecycle management, enabling public authorities and stakeholders to create, analyse, evaluate and optimize policies based on the “fresh” data, the information that can be continuously collected by citizens and other sensors. These technologies will be further examined and exploited within the framework of PolicyCloud so as to provide an integrated web-based environment to fulfil the requirements of advanced policy lifecycle management.

8 Background Technologies

The development of the PolicyCLOUD platform will be based on already existed baseline technologies that the partners of the consortium are bringing to the project as the background and they plan to further develop them in order to fulfil the requirements of the platform, as they have been listed in the previous subsections. The following table contains a list of those indicative baseline technologies that are considered to be exploited by the PolicyCLOUD platform. It is worth to mention that at this phase of the project, it is not certain if the list is exhaustive or it can be further extended in the next iterations of this deliverable.

Technology Name	Technology Description	Advancements / Usage
LeanXcale DataStore	A highly scalable relational database management system, that ensures transactional semantics and provides an efficient manner to deal with highly ingestion rates. Additionally, it offers an parallel analytical query engine that makes it possible to retrieve data, while data is being ingested, thus, can be considered as an HTAP database, which allows for combining analytical and operational workload on the same data. It also can be easily extended to provide polyglot support.	In the scope of PolicyCLOUD, the internal query engine of the datastore will be extended in order to achieve a greater level of parallelism, and thus being able to serve analytical queries much more efficiently. Moreover, its internal polyglot support will be further extended in order to provide a common manner to access and join data storing in other stores, contributing to the fusion of the data.
Capturean Tool	Solution for SN monitoring. Available for Twitter mostly. Provides sentiment analysis and other SN metrics for specific listening channels (topics) defined by the customers. The solution provides a dashboard to visualize the results and a REST API to access to it programmatically	Sentiment Analysis over RRSS (mainly Twitter)
IBM Data Skipping and Smart Layout library	Library for Apache Spark for creating and using metadata indices that optimize SQL-based analytics, adjust the data layout for analytics optimization.	Dramatic performance improvement for SQL-based analytics over big data in object storage
Policy Development Toolkit	A framework for creating and evaluating policies related with the healthcare section	In the scope of PolicyCLOUD this asset will be further extended in order to allow the creation of general scope policies
Interoperability mechanism	Data interoperability realized through a process of identifying the structural and semantic similarity of domain-specific knowledge to turn the datasets into interoperable domain-agnostic ones.	In PolicyCLOUD the data interoperability mechanism will be extended towards different types of datasets and formats emerging from the identified PolicyCLOUD underlying data sources.
Sources reliability tool	Sources reliability tool for mapping heterogeneous IoT devices into specific levels of trustfulness, thus estimating the overall reliability of each data source.	In PolicyCLOUD the sources reliability tool will be extended for estimating the reliability of all the available different types of data sources, and thus keeping into the platform for further analysis only the data that comes from only reliable sources.

TABLE 169: BASELINE TECHNOLOGIES

Moreover, the majority of the partners of the consortium have great experience in participating in other on-going European and National research projects, whose topic of interest are relevant to the PolicyCLOUD. Due to this, outcomes and assets that have been developed in those research projects are candidates to be part of the platform or to further extend their functionalities in order to address PolicyCLOUD specific requirements. The following table contains a non-exhaustive list of other projects that might be useful in the development of the platform, along with information on how the latter can benefit from their use.

Project	Relevant Result	Advancement in PolicyCLOUD
CrowdHEALTH	• We will use the Policy Development Toolkit that was firstly introduced in the scope of the CrowdHEALTH project. This is a framework for Analytics Tools registration into the Back-end and communication with the Policy Development Toolkit front-end. • Extended model of interoperable health data, and mechanism for estimating data sources' reliability.	• UI Dashboard for policy development. Personal workspace for policy makers. Parameter selection functionality for the invocation of Analytics. User notification for Analytics results. On-line help provision to the policymakers. • Utilize the interoperability model, and the data sources reliability calculation metrics for additional datasets/cases identified in the context of PolicyCLOUD.
CoherentPaaS	• A common query language that can be used from a polystore in order to retrieve data resigns in different and heterogeneous datastores. • A polyglot query engine that can execute queries addressing different datastores	The outcomes of CoherentPaaS will be used in order to implement the data fusion of the platform. They will be further extended to achieve the maturity required for the needs of the project, as the current state is a prototype. Further extensions will be made to support the different use cases
BigDataStack	• Data layout extensions for the IBM Object Store, in order to accelerate analytical queries. • A seamless analytical framework that combines the benefits of an operational database and data warehouse, moving historical data from the former o the latter	• More accurate data layout. • Support for all SQL data operations from the seamless analytical framework

TABLE 170: RELEVANT RESEARCH PROJECTS

9 Conclusion

This document firstly summarized the methodology that was agreed in the scope of the T2.1 of the project for collecting the user and technical requirements of the project. Based on this methodology, a list of concrete scenarios for each of the use case was specified, along with the initial version of their relevant user requirements. What is more, the technical partners of the consortium also provided the initial set of technical requirements as they were foreseen at this starting phase of the project. Additionally, it provided the state-of-the-art analysis of the base technology sectors that the PolicyCLOUD project is involved, and could possibly exploit, along with a list of baseline technological tools and solutions that are planned to be incorporated in the overall platform. At this initial phase of the project, the outcomes of this deliverable have created valuable input for the progress of the task that is related to the design of the overall architecture of the platform. Moreover, a brief analysis of the various stakeholder roles and their relevant business goals has been conducted that will assist the tasks related with the market analysis and the identification of business potentials for the platform.

This is the first of a series of versions that are planned to be released through the project. On M12, a second version will update the current list of the user and technical requirements, taking into considerations that the use case will be more mature and their relevant scenarios will be better defined. At that point, the overall architecture will need to be further refined and extended in order to cover more advanced scenarios that were not taken into account at this early phase. Finally, a third version is planned to be delivered on M22, in order to cover or remaining aspects and to correct potential erroneous decisions or unnecessary requirements that might have been identified earlier, so that it can drive the final definition of the requirements that will drive the overall architecture of the project, as the latter will be heading towards to its conclusion.

References

[1] A Survey of Data Fusion in Smart City Applications, Lau at al. ELSEVIER INFORMATION FUSION 2019

[2] Real-time pricing by data fusion on networks," Izumi at al, IEEE Transactions on Industrial Informatics, vol. 14, no. 3, 2018

[3] Iterative channel estimation using lse and sparse message passing for mmwave mimo systems, Huang at al., IEEE Transactions on Signal Processing, vol. 67, no. 1, 2019.

[4] Feature learning and analysis for cleanliness classification in restrooms, Jayasinghe at al, IEEE Access, vol. 7, 2019

[5] Understanding the lifestyle of older population: Mobile crowdsensing approach, Marakkalage at al., IEEE Transactions on Computational Social Systems, vol. 6, no. 1, 2019

[6] Harnessing multi-source data about public sentiments and activities for informed design, You at al., IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 2, 2018

[7] Internet of things for green building management: Disruptive innovations through low-cost sensor technology and artificial intelligence, Tushar at al., IEEE Signal Processing Magazine, vol. 35, no. 5, 2018

[8] A. Solanas et al., "Smart Health: A Context-Aware Health Paradigm within Smart Cities," IEEE Comm. Mag., vol. 52, no. 8, 2014, pp. 74–81, 2014.

[9] Melnik S., Decker S., "A Layered Approach to Information Modeling and Interoperability on the Web," Proc. ECDL 2000 Workshop Semantic Web, http://infolab.stanford.edu/~melnik/pub/sw00/sw00.pdf, 2000.

[10] New European Interoperability Framework, https://ec.europa.eu/isa2/sites/isa/files/eif_brochure_final.pdf , Accessed 2 Mar 2020.

[11] Hellmann S., Lehmann J., Auer S., Brümmer M., Integrating NLP using Linked Data,12th International Semantic Web Conference, pp. 21-25, 2013.

[12] Fernandez, R.C., Mansour, E., Qahtan, A.A., et al., Seeping semantics: Linking datasets using word embeddings for data discovery, In: 2018 IEEE 34th International Conference on Data

[13] Farias T. M., Stockinger K., Dessimo C, VoIDext: Vocabulary and Patterns for Enhancing Interoperable Datasets with Virtual Links, 2019.

[14] Binding C., May K., Tudhope D., Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction Via the CIDOC CRM. In: Christensen-Dalsgaard B., Castelli D., Ammitzbøll Jurik B., Lippincott J. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2008. Lecture Notes in Computer Science, vol 5173. Springer, Berlin, Heidelberg (2008)

[15] Colpaert P., Van Compernolle M., et. al., Quantifying the Interoperability of Open Government Datasets, DOI: 10.1109/MC.2014.296, 2014.

[16] JSON-LD - JSON for Linking Data, http://json-ld.org , Accessed 5 Mar 2020.

[17] Xin, J., Afrasiabi, C., Lelong, S., Adesara, J., Tsueng, G., Su, A. I., Wu, C., Cross-linking BioThings APIs through JSON-LD to facilitate knowledge exploration, BMC bioinformatics, 19(1), 30. https://doi.org/10.1186/s12859-018-2041-5 ,(2018)

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

[19] Wikipedia (17 March 2020). Comparison of JavaScript charting libraries, https://en.wikipedia.org/wiki/Comparison_of_JavaScript_charting_libraries, retrieved 2020-03-25)

[20] Barot, Saurabh (2020). Top JavaScript Chart Libraries to Use in 2020 for Better Data Visualization, https://aglowiditsolutions.com/blog/top-javascript-chart-libraries/, retrieved 2020-03-25)

[21] Borovikov,Ruslan (April 27th 2019). Top 10 JavaScript Charting Libraries for Every Data Visualization Need, https://hackernoon.com/10-javascript-charting-libraries-data-visualization-b77523d23372, retrieved 2020-03-26)

[22] Puszynski,Arthur (15 May 2019). These are the best JavaScript chart libraries for 2019, https://www.freecodecamp.org/news/these-are-the-best-javascript-chart-libraries-for-2019-29782f5e1dc2/, retrieved 2020-03-26)

[23] Shealy,Matt (21 Nov 2019). Best Data Visualization Tools for 2020 Reviewed, https://readwrite.com/2019/11/21/best-data-visualization-tools-for-2020-reviewed/, retrieved 2020-03-26)

[24] Goel,Aman (29 Feb, 2020). 10 Best JavaScript Frameworks to Use in 2020, https://hackr.io/blog/best-javascript-frameworks, retrieved 2020-03-26)

[25] Sviatoslav A. (JAN 08, 2020). The Best JS Frameworks for Front End, https://rubygarage.org/blog/best-javascript-frameworks-for-front-end, retrieved 2020-03-26)

[26] Apache Drill – Schema-free SQL Query Engine for Hadoop, NoSQL and Cloud Storage, https://drill.apache.org/

[27] Apache Impala, http://impala.apache.org/

[28] M. Armbrust, R. Xin, C. Lian, Y. Huai, D. Liu, J. Bradley, X. Meng, T. Kaftan, M. Frank-lin, A. Ghodsi, M. Zaharia, "Spark SQL: relational data processing in Spark", in ACM SIGMOD, 2015, pp. 1383-1394.

[29] F. Bugiotti, D. Bursztyn, A. Deutsch, I. Ileana, I. Manolescu, "Invisible glue: scalable self-tuning multi-stores", in Conference on Innovative Data Systems Research (CIDR), 2015.

[30] S. Dasgupta, K. Coakley, A. Gupta, "Analytics-driven data ingestion and derivation in the AWESOME polystore", in IEEE International Conference on Big Data, 2016, pp. 2555-2564.

[31] J. Duggan, A. J. Elmore, M. Stonebraker, M. Balazinska, B. Howe, J. Kepner, S. Madden, D. Maier, T. Mattson, S. Zdonik, "The BigDAWG polystore system", SIGMOD Record, vol. 44, no. 2, pp. 11-16, 2015.

[32] V. Gadepally, P. Chen, J. Duggan, A. J. Elmore, B. Haynes, J. Kepner, S. Madden, T. Mattson, M. Stonebraker, "The BigDawg polystore system and architecture", in IEEE High Performance Extreme Computing Conference (HPEC), 2016, pp. 1-6.

[33] Y. Khan, A. Zimmermann, A. Jha, D. Rebholz-Schuhmann, R. Sahay, "Querying web pol-ystores", in IEEE International Conference on Big Data, 2017.

[34] Z. Minpeng, R. Tore, "Querying combined cloud-based and relational databases", in Int. Conf. on Cloud and Service Computing (CSC), 2011, pp. 330-335.

[35] K. W. Ong, Y. Papakonstantinou, and R. Vernoux, "The SQL++ semi-structured data model and query language: a capabilities survey of SQL-on-Hadoop, NoSQL and NewSQL databases", CoRR, abs/1405.3631, 2014.

[36] T. Özsu, P. Valduriez, Principles of Distributed Database Systems, 3rd ed. Springer, 2011, 850 pages.

[37] Presto – Distributed Query Engine for Big Data, https://prestodb.io/

[38] A. Simitsis, K. Wilkinson, M. Castellanos, U. Dayal, "Optimizing analytic data flows for multiple execution engines", in ACM SIGMOD, 2012, pp. 829-840.

[39] A. Tomasic, L. Raschid, P. Valduriez, "Scaling access to heterogeneous data sources with DISCO", IEEE Trans. On Knowledge and Data Engineering, vol. 10, pp. 808-823, 1998.

[40] J. Wang, T. Baker, M. Balazinska, D. Halperin, B. Haynes, B. Howe, D. Hutchison, S. Jain, R. Maas, P. Mehta, D. Moritz, B. Myers, J. Ortiz, D. Suciu, A. Whitaker, S. Xu, "The Myria big data management and analytics system and cloud service", in Conference on Innovative Data Systems Research (CIDR), 2017.

[41] Banks, G. (2009), Evidence-based policy making: What is it? How do we get it? (ANU Public Lecture Series, presented by ANZSOG, 4 February), Productivity Commission, Canberra.

[42] Howlett, M. & Ramesh, M. (2003), Studying public policy: Policy cycles and policy subsystems. Toronto, ON: Oxford University Press Canada.

[43] Big Policy Canvas. D5.2 Roadmap for Future Research Directions – Final Version. Big Policy Canvas, Francesco Mureddu. 2019.

[44] Shroff ZC, Javadi D, Gilson L, Kang R, Ghaffar A. Institutional capacity to generate and use evidence in LMICs: current state and opportunities for HPSR. Health Res Policy Syst. 2017 Nov 9;15(1):94.

[45] Zardo P, Collie A. Type, frequency and purpose of information used to inform public health policy and program decision-making. BMC Public Health. 2015 Apr 15;15:381.

[46] O'Donoughue Jenkins L, Kelly PM, Cherbuin N, Anstey KJ. Evaluating and Using Observational Evidence: The Contrasting Views of Policy Makers and Epidemiologists. Front Public Health. 2016;4:267.

[47] Lee BY, Ferguson MC, Hertenstein DL, Adam A, Zenkov E, Wang PI, et al. Simulating the Impact of Sugar-Sweetened Beverage Warning Labels in Three Cities. Am J Prev Med. 2018 Feb;54(2):197–204.

[48] Powell-Wiley TM, Wong MS, Adu-Brimpong J, Brown ST, Hertenstein DL, Zenkov E, et al. Simulating the Impact of Crime on African-American Women's Physical Activity and Obesity. Obes Silver Spring Md. 2017 Dec;25(12):2149–55.

[49] Atkinson J-A, Knowles D, Wiggers J, Livingston M, Room R, Prodan A, et al. Harnessing advances in computer simulation to inform policy and planning to reduce alcohol-related harms. Int J Public Health. 2018;63(4):537–46.

[50] Freebairn L, Atkinson J-A, Kelly PM, McDonnell G, Rychetnik L. Decision makers' experience of participatory dynamic simulation modelling: methods for public health policy. BMC Med Inform Decis Mak [Internet]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6291959/

[51] Katrakazas P, et al. Public Health Policy Decision Models (PHPDM) v1, Deliverable D3.1 to the EVOTION-727521 Project funded by the European Union, ICCS

[52] Kyriazis D, Autexier S, Brondino I, Boniface M, Donat L, Engen V, et al. CrowdHEALTH: Holistic Health Records and Big Data Analytics for Health Policy Making and Personalized Health. Stud Health Technol Inform. 2017;238:19–23

[53] Moutselos K, Kyriazis D, Diamantopoulou V, Maglogiannis I. Trustworthy data processing for health analytics tasks. In: 2018 IEEE International Conference on Big Data (Big Data). 2018. p. 3774–9.

[54] Moutselos K, Maglogiannis I. Evidence-based Public Health Policy Models Development and Evaluation using Big Data Analytics and Web Technologies. Medical Archives. 2020; 74(1): 47-53.

[55] Kamateri E, Panopoulou E, Tambouris E, Tarabanis K, Ojo A, Lee D, et al. A Comparative Analysis of Tools and Technologies for Policy Making. In: Janssen M, Wimmer MA, Deljoo A, editors. Policy Practice and Digital Science: Integrating Complex Systems, Social Simulation and Public Administration in Policy Research. Cham: Springer International Publishing; 2015. p. 125–56.

[56] P. J. Giabbanelli and M. Baniukiewicz, "Navigating Complex Systems for Policymaking Using Simple Software Tools," in Advanced Data Analytics in Health, Springer, Cham, 2018, pp. 21–40.