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

D5.8 METHODS FOR CLOUD USE BY PUBLIC AUTHORITIES M32

Dissemination Level	PU
Due Date of Deliverable	31/08/2022, Month 32
Actual Submission Date	01/09/2022
Work Package	WP5
Task	T5.1
Type	Report
Approval Status	
Version	v1.0
Number of Pages	p.1 – p.79

Abstract: Deliverable D5.8. is the second expanded iteration of Deliverable D5.1., submitted in August 2021. This report showcases how the methodology and the approach for creating policies using Cloud infrastructures and Big Data technologies proposed in D5.1. is applied, specifying how to use the value of different mechanisms during the entire lifecycle of the policy in the scope of the Cloud environments.

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
0.1	23/08/2021	Submitted version of D 5.1	Atos
0.2	25/05/2022	Initial Team Discussion on D 5.8	Ana Georgieva (SOF) Dobrinka Blagova (SOF) Metodiyka Tarlyovska (SOF) Martin Atanasov (SOF)
0.3	01/06/2022	Initial ToC and Assignments	Dessislava Petrova (SOF) Konstantin Antonov (SOF) Metodiyka Tarlyovska (SOF) Martin Atanasov (SOF)
0.4	15/06/2022	ToC and Assignments – Sofia Team Methodology for Policy Making in Cloud Environment	Dessislava Petrova (SOF) Konstantin Antonov (SOF) Metodiyka Tarlyovska (SOF) Martin Atanasov (SOF)
0.5	15/07/2022	ToC and Assignments - Consortium	Dessislava Petrova (SOF) Martin Atanasov (SOF) Metodiyka Tarlyovska (SOF)
0.6	22/07/2022	Section 3	Rafael del Hoyo (SARGA) Vega Rodrigálvarez (SARGA)
0.7	18/08/2022	Consolidated Draft Ready for Internal Review	Martin Atanasov (SOF) Metodiyka Tarlyovska (SOF)
0.8	21/08/2022	Internal Review	Petya Nikolova (OKYS)
0.9	21/08/2022	Internal Review	Sara Diez Minguez (ATOS)
0.10	21/08/2022	Internal Review	George Manias (UPRC)
0.11	21/08/2022	Internal Review	Kostas Nasias (OKYS)
0.12	22/08/2022	Addressing Review Comments	Martin Atanasov (SOF) Metodiyka Tarlyovska (SOF)
0.13	31/08/2022	Quality Check Performed and Changes Addressed	Metodiyka Tarlyovska (SOF)
0.14	01/09/2022	Final version ready for submission	Metodiyka Tarlyovska (SOF)
1.0	01/09/2022	Submitted version	Atos

Author List

Organisation	Name
SOF	Ana Georgieva Dessislava Petrova Dobrinka Blagova Konstantin Antonov Metodiyka Tarlyovska Martin Atanasov
ATOS	Ana Luiza Pontual Maria Angeles Sanguino Gonzalez Miquel Mila Prat
IBM	Yosef Moatti
ICTLC	Alice Audino Alberto Bettiol Martim Taborda Barata
LON	Ben Williams
MAG	Armend Duzha Samuele Baroni
OKS	Petya Nikolova Kostas Nasias
SARGA	Javier Sancho
UPRC	Argyro Mavrogiorgou George Manias Thanos Kiourtis
UBI	Giannis Ledakis

Abbreviations and Acronyms

Abbreviation/Acronym	Definition
AI	Artificial Intelligence
API	Application Programming Interface
BDA	Big Data Analytics
BPaaS	Business Process-as-a-Service
CaaS	Container-as-a-Service
CIA principle	Confidentiality, Integrity, Availability principle in Big Data
CoE	Centre of Excellence
DaaS	Data-as-a-Service
DAEU	Bulgarian State E-Government Agency
DESI	Digital Economy and Society Index
DM	Data Marketplace
DPIA	Data Protection Impact Assessment
EC	European Commission
EOSC	European Open Science Cloud
FaaS	Function-as-a-Service
FAIR principle	Findability, Accessibility, Interoperability and Reusability principle
GDPR	General Data Protection Regulation
IaaS	Infrastructure-as-a-Service
IoT	Internet of Things
KPI	Key Performance Indicator
MTITC	Ministry of Transport, Information Technology and Communications
NaaS	Network-as-a-Service
NCSC	(UK) National Cyber Security Centre
NGO	Non-Governmental Organisation
NIST	(American) National Institute of Standards and Technology
NKOS	Networked Knowledge Organization Systems
NLP	Natural Language Processing
OGL	Open Government License
OS	Operation Systema
PA	Public Authority
PaaS	Platform-as-a-Service
PDT	Policy Development Toolkit
PKI	Public Key Infrastructure
PP	Public Policy
SaaS	Software-as-a-Service
SMEs	Small and Medium-Sized Enterprises
SOA	Service-Oriented Architecture
UI	User Interface
UX	User Experience
VPN	Virtual private network
WP	Work Package

Contents

List of Tables

List of Figures

Executive Summary

This document is the sixth of the series of deliverables developed under WP5, which aims to realize the overall goal of policy modelling, development and implementation, enabling the analysis of data to be directly exploited in different phases of the policy lifecycle.

This deliverable has been submitted in September 2022 (M33 of the PolicyCLOUD project). It is the second expanded iteration of the deliverable D5.1 'Methods for Cloud Use by Public Authorities' (submitted in August 2021). The main purpose of Deliverable D5.8 is to illustrate how the methodology presented in D5.1, specifying how to exploit the value of different mechanisms (e.g., data modelling, semantic representation and interoperability, analytics, etc.) during the policy lifecycle (from modelling to implementation) in the scope of Cloud environments, as well as the set of proposed adaptive techniques for creating policies using cloud infrastructures and Big Data technologies, can be beneficial for the activities of various institutions with regards to their data-based decision- and policy- making processes.

Deliverable D5.8 is aimed at the PolicyCLOUD partners and participants, a wide range of stakeholders, involved in the topics of data, Big Data and Cloud computing, as well as the wider community of public administrations (PAs), decision-makers and citizens interested in taking advantage of the capabilities and benefits of Cloud technologies and Big Data in their policy-making. It follows the structure of Deliverable D5.1. and includes additional information and further advancements and contributions from other tasks.

The number of concrete examples and success stories related to the use and utilisation of the Cloud and Big Data in the public sector in different countries (Section 2.2.) has been expanded. In addition to the four PolicyCLOUD pilot use cases (namely Italy (Lombardy Region), Spain (Autonomous Community of Aragon), Bulgaria (Sofia), and the UK (Camden Town, London)), the document now includes good practices from the City of Barcelona (Spain) and Germany – as rapidly developing digital, innovative and smart city hubs, with the aim to raise further awareness and serve as an inspiration for the representatives of the public sector and a wider range of stakeholders regarding the capabilities and the value of the Cloud and Big Data in the public sector. A new section dedicated to data space ecosystems in the public sector has been added as well (Section 3.6).

The *Methodology for Policy Making in Cloud Environment*, initially proposed in D5.1. (Section 4) has been expanded as well, and Sofia's use case during the PolicyCLOUD project is used as a concrete practical example for the methodology validation (Section 4.3). To present the application of the methodology within the context of the PolicyCLOUD, all the steps that users need to take from creating a specific policy, to saving the policy in the Cloud, and then being able to put it into practice and receiving analyses that they can use in their work are illustrated.

1 Introduction

1.1 Purpose of the document

This document is the enriched version of the deliverable D5.1 'Methods for Cloud Use by Public Authorities', covering task T5.1, and the sixth deliverable of WP5. It targets both PolicyCLOUD participants and the wider community of public administrations (PAs) and decision-makers who are interested in reaping the benefits of the Cloud and Big Data in policy-making.

During the one year between the two iterations of this series of deliverables (August 2021 – August 2022), the PolicyCLOUD Consortium has managed to develop a comprehensive data processing process for the analysis of data and group them properly in order to obtain an improved result for the analyses on the basis of which to develop policies.

The main purpose of D5.8 is to upgrade the approach and methodology for policy-making using Cloud infrastructures and Big Data technologies initially suggested in D5.1. which: 1) specifies how to use the value of different mechanisms (e.g. data modelling, semantic presentation and interoperability, visualisation, analysis, etc.); 2) is applicable throughout the life cycle of cloud policy; and 3) proposes a set of adaptive techniques for using Cloud-based policy-making.

As already mentioned in the Executive Summary, Deliverable D5.8 follows the structure of Deliverable D5.1. and includes additional information and further advancements and contributions from other tasks. The document presents concrete examples based on working with the PolicyCLOUD platform in different use cases to illustrate the ways of creating policies, as well as what types of analytics each user can get by using the PolicyCLOUD tools to improve the urban environment. This report is enriched with additional practices for the use of Cloud infrastructures from different EU countries for the purpose of processing databases and their reuse.

1.2 Structure of the document

The rest of the document is structured as follows. **Section 2** describes examples of the use of Cloud and Big Data in the public sector, including:

- definition of Cloud computing and Big Data and an overview of their value and benefits in the context of both the public sector and the PolicyCLOUD project (Section 2.1);
- the use of Cloud architectures and models and Big Data in the public sector with concrete examples, success stories and benchmarks from countries around the world, starting with the four PolicyCLOUD pilot use cases: Italy (Lombardy Region), Spain (Autonomous Community of Aragon) Barcelona, Bulgaria (Sofia), the UK (Camden Town, London), and Germany (Berlin) (Section 2.2).

Section 3 presents an overview of the data lifecycle of public policies, including the key methods, approaches and technologies and their use in the PolicyCLOUD context, with special focus on the ethical, legal, regulatory, and societal considerations.

Section 4 focuses on the methodology for policy making in Cloud environments. The proposed methodology adapts Kaizen methodology and the concept of the digital twins and has 2 main inputs: knowledge and data. Moreover, Section 4 focuses on the methodology elaboration/development (Section 4.1); prerequisites for usage and applicability (Section 4.2); and methodology validation (Section 4.3).

Finally, **Section 5** provides the conclusion of the document.

2 Cloud and Big Data in the Public Sector

The Cloud and Big Data are among the key factors enabling the development, digitalisation and modernisation of the public sector, its more effective and efficient decision- and policy-making, and better collaboration within various groups of stakeholders – the private sector, the academia, NGOs, the civil society, the innovative and entrepreneurial ecosystem, etc. Their value and benefits can be experienced both via their support of more efficient processes, coordination and use of resources, and the encouragement of the development of innovative solutions and products.

This section includes:

- a brief overview of the Cloud and Big Data and their usage in the public sector;
- the key value and benefits of the Cloud and Big Data for the public sector, including in the context of the PolicyCLOUD project;
- examples of good practices related to the use and utilisation of the Cloud and Big Data in the public sector in different countries, starting with the PolicyCLOUD pilot use cases: Italy (Lombardy Region), Spain (Autonomous Community of Aragon), Bulgaria (Sofia), and the UK (Camden Town, London), as well as from the City of Barcelona (Spain) and Germany, and in particular the City of Berlin – as rapidly developing digital, innovative and smart city hubs.

2.1 Definition, Value and Benefits

2.1.1 Definition, Value and Benefits of the Cloud in the Public Sector

Cloud computing and Cloud solutions have been increasingly important in the EU and worldwide, drawing the attention of the public and private sector, the researchers, and the civil society, with their adoption on the rise and focus on enabling increasing cost, process and resource efficiency and flexibility. The Cloud is a key technology supporting the development of the global economy. According to the World Economic Forum, in addition to reducing operational costs, the Cloud has become the basis for radical business innovation, new business models, and significant improvements in the effectiveness of anyone using IT – governments, companies, individuals, academia, etc. [1]

2.1.1.1 The Cloud: Definition and Essential Characteristics

The five essential characteristics of the Cloud, as described by NIST, are being enabled by the Cloud infrastructure and showcase the value and benefits of the Cloud [2]:

1. On-demand self-service: automatic and unilateral access to capabilities (e.g. network storage or server time, etc.), without the need of help or interaction with each service provider.
2. Broad network access: availability of a wide variety of capabilities over the network which can be accessed through different client platforms (e.g., mobile phones, tablets, laptops, etc.)
3. Resource pooling: the offered capabilities (physical or virtual, e.g. storage, server memory, network processing, security, applications, and services, etc.) are pooled to be used by multiple consumers

(multi-tenant model) and are being assigned and reassigned in a dynamic demand-based manner. The user generally doesn't know the exact location of the resources and has no control over them, but can specify, for example, a concrete country, data centre, etc.

4. Rapid elasticity: resources are being provided elastically, sometimes automatically, to be rapidly scalable outward or inward depending on the demand. To the user, the capabilities often seem to be unlimited and accessible at any time and quantity.
5. Measured service: Cloud systems can leverage their metering capability (usually on a pay-per-use or charge-per-use basis) depending on the type of services provided to automatically control and optimise resource use, which can be monitored, controlled, and reported to provide transparency for both the provider and the users.

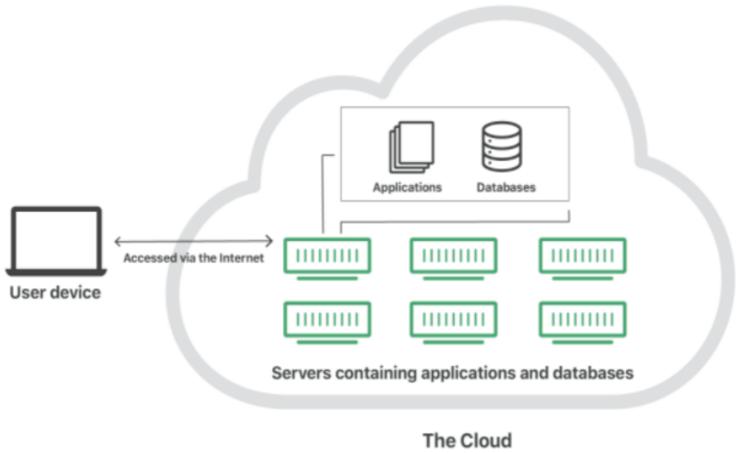


FIGURE 1: THE CLOUD. SOURCE CLOUDFLARE (2021) [3]

Cloud Service (or Architecture) Models. Formerly, there have been three main models of Cloud computing: SaaS – Software-as-a-Service, PaaS – Platform-as-a-Service, and IaaS – Infrastructure-as-a-Service. However, other models have emerged, among which FaaS or Function-as-a-Service, also known as serverless computing, and CaaS or Container-as-a-Service:

- SaaS: users access the capabilities and applications running on a Cloud infrastructure from different types of devices via a thin client interface or a program interface. Generally, they have no control over the underlying infrastructure apart from some application settings.
- PaaS: users can deploy different applications (acquired or developed by them) to the Cloud infrastructure, developed through services and tools supported by the provider. As in SaaS, the consumer has no control over the underlying infrastructure, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.
- IaaS: users have access to processing, storage, networks, etc. computing resources, and can deploy and run arbitrary software (e.g. operating systems and apps). They have no control over the underlying infrastructure, but have control over operating systems, storage, and deployed applications; possibly, limited control of select networking components (e.g., host firewalls).
- FaaS: allows users to execute code when needed without the complex infrastructure typically associated with building and launching microservice applications [4]. It breaks Cloud applications down into even smaller components that only run when they're needed. FaaS still runs on servers

but is known as 'serverless computing' as: 1) it does not run on dedicated machines, and 2) the users building the applications do not have to manage any servers [3].

- CaaS: a type of container-based virtualization, where container engines, orchestration and the underlying compute resources are delivered to users as services, which can be deployed on-premises or in a Cloud [5]. The Cloud provider offers the orchestration platform, where the containers are deployed and managed, and through the orchestration the key IT functions are being automated. Containers allow the rapid development and delivery of Cloud-native applications that can run anywhere, and users can buy only the capabilities they need [6].

In addition, other types of service models are being currently provided, such as:

- Business Process as a Service (BPaaS): emerges from SaaS and implies the provision of Cloud-sourced fully configurable process outsourcing to various customers.
- Data as a Service (DaaS): the provision of Cloud-based data-related services (e.g. storage, processing, analytics, etc.).
- Network as a Service (NaaS): instead of using their own network infrastructure, consumers can rent Cloud-based networking services from a Cloud provider.

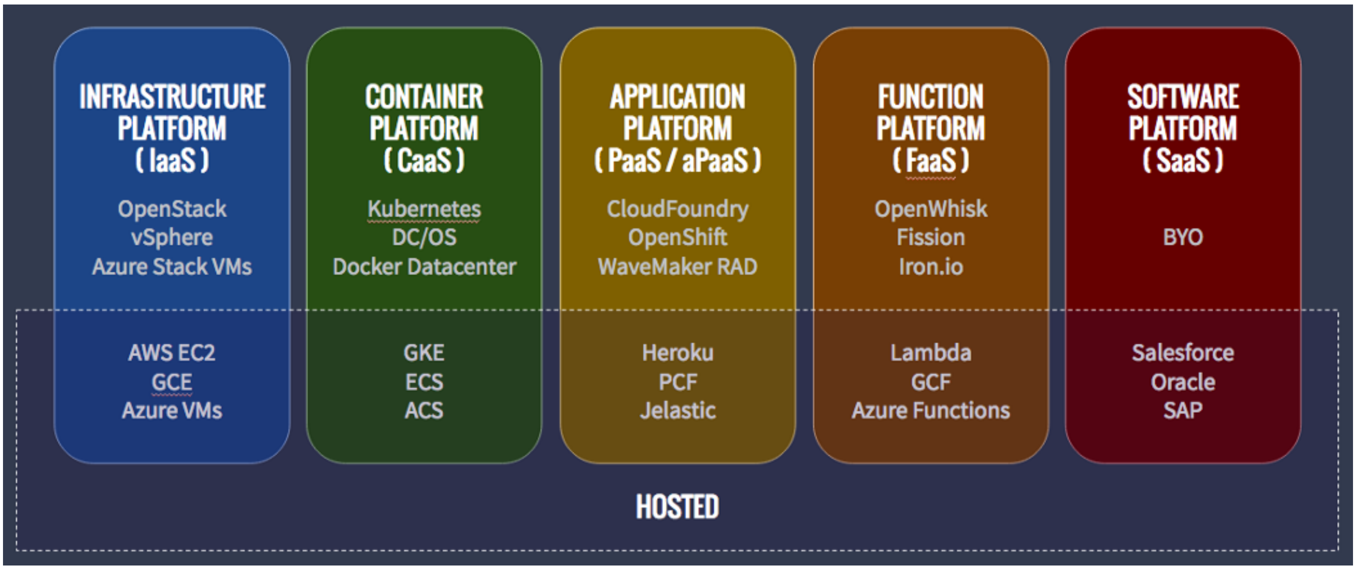


FIGURE 2: THE MAIN CLOUD SERVICE MODELS. SOURCE MEDIUM [5]

Cloud Deployment Models: There are four main types of Cloud deployment models (as defined by NIST, 2011): private cloud, community cloud, public cloud and hybrid cloud, with a fifth, additional type known as Multicloud:

- Private cloud: provided to a single organisation consisting of multiple consumers for its exclusive use, it may be owned, managed, and operated by the organisation, a third party, or some combination of them, and may exist on or off premises. It is more secure than public clouds and allows cost savings and better resource utilisation.

- Community cloud: provided to a specific community of users from organisations with shared concerns (e.g., mission, compliance considerations, etc.) for their exclusive use. As with the private cloud, it may be owned, managed, and operated by one or more of the organisations in the community, a third party, or some combination of them, and may exist on or off premises.
- Public cloud: provided for open use to the general public, it may be owned, managed, and operated by the provider offering the Cloud service (e.g. a business, academic, or government organisation, or some combination of them), and exists on his/her premises. It is hosted online, and the users can access it anytime as long as they have internet connection.
- Hybrid cloud: more complex than the other deployment models, it is a combination of two or more separate Cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardised or proprietary technology that enables data and application portability (e.g., Cloud bursting for load balancing between clouds).
- Multicloud: a fifth, additional type of Cloud deployment that involves using multiple public clouds. An organisation with a multicloud deployment rents virtual servers and services from several external vendors. Multicloud deployments can also be hybrid cloud, and vice versa [3].

Essential Characteristics of the Cloud	
5 Essential Characteristics of the Cloud	On-demand self-service, broad network access, resource pooling, rapid elasticity, measured service
Key Cloud service (architecture) models	SaaS - Software as a Service; PaaS - Platform as a Service; IaaS - Infrastructure as a Service; FaaS - Function-as-a-Service
Key Cloud deployment models	Private cloud; Community cloud; Public cloud; Hybrid cloud; Multicloud

TABLE 1: ESSENTIAL CHARACTERISTICS OF THE CLOUD

2.1.1.2 Value and Benefits of the Cloud for the Public Sector

We live in an era of rapid and unprecedented change, which has only been amplified by the COVID-19 crisis. The need for new technologies and businesses through which society strives to become greener, more innovative, digital, sustainable, transparent and efficient. The same applies to the public sector, which aims to provide the best services and quality of life for citizens and businesses.

The need for more Cloud spaces to provide an opportunity for people from all areas to carry out their duties using the capabilities of Cloud technologies has been proven. Public and private organisations are using the Cloud as an important tool to facilitate their digital transformation. Cloud computing and virtualization (which makes physical location irrelevant for resource allocation and service delivery) offer new opportunities for organisations to address long-term and short-term challenges, allowing them to reduce risks and use technology for competitive advantage [7], [8]. A disruptive technology, the Cloud facilitates the development of innovative solutions, services and models in the public and private sectors, academia, the entrepreneurial ecosystem, civil society and more. Through virtualization, clouds address a larger number of users with different needs using the same shared resources, providing an alternative to clusters, networks, and supercomputers [9].

Cloud computing increases time and operational efficiency; facilitates the development of new methods of using and applying technologies, including green IT; modernise the models by which the sector develops and provides its services; strives for better information protection, faster and trouble-free service provision, accountability and transparency; and facilitates the engagement of a wide variety of stakeholders in new ways, including collaboration and co-production of knowledge [10]. It also enables cost optimization, and the public sector has always sought to provide better public services and a high quality of life within its limited budgets. Organisations not only rely on cCloud services to improve their internal efficiency, but also to target more strategic capabilities with the key objective of building external capacity for increased collaboration with external partners [11]. The benefits of the public sector cloud can be compared to the benefits gained through the use of utilities, where consolidating ICT assets often makes more sense than developing them for use at individual sites or as separate systems. As public organisations are used to sharing data and resources, Cloud computing can be seen as the natural next step in their approach to shared services [8].

For the purposes of this deliverable, we propose an exhaustive list of key advantages and benefits of Cloud solutions for the public sector and the society as a whole (with a special focus on the public sector) that fall into 4 main categories: 1) Increased flexibility and adaptability; 2) Cost-benefit optimisation and control; 3) Increased efficiency, better services and collaboration; 4) Strategic value: facilitating innovation and business transformation. The list and the proposed categorisation are based on a number of resources, and mainly on Macias&Thomas [8]; IBM [12], [13]; Müller et al. [10]; Cox [14].

Category	Value/Benefits/Advantages
1. Increased flexibility and adaptability: Cloud resources are dynamic and available at all times. The users can scale and customise the services to fit their needs and access them from anywhere via the internet.	• Accessibility: dynamic availability of Cloud-based solutions and data, accessible from various internet-connected devices • Simple scalability on demand • Storage options: choice of a deployment model • Control: choosing the level of control with the different as-a-service options (SaaS, PaaS, IaaS, etc.) • Easily shared resources: across organisations/cities/regions/countries • Variety of tools, features and services to fit specific needs with convenient, on-demand access to a common pool of configurable computing resources • Elastic services: easier service expansion via shared pools of resources or implementing pre-packaged capabilities • Effective risk mitigation: various methods and technologies to ensure reliability and that data and privacy are safeguarded • Security features: e.g. private cloud, encryption, API keys, etc. In general, data in a Cloud can be secured with as much confidence as in

Category	Value/Benefits/Advantages
	a closed network, if the system is equipped with appropriate protective measures and is well maintained. ● Agility and adaptability: quicker reactions to changes, more smooth and efficient innovation process, less burdensome and expensive processes of change. Use of virtual integration to allocate resources to specific processes whenever needed
2. Cost-benefit optimisation and control: the public sector aims to cut costs without cutting on quality or critical services. The Cloud allows for cost-benefit optimisation as users don't have to take care of maintenance and underlying infrastructure costs	● Savings on cost: e.g. potential real estate costs for a data centre, power and cooling costs, etc. ● Cost reduction and variabilisation via specialisation and economy of scale, increased process efficiency and reduced redundancy (e.g. due to the pay-per-use nature of the Cloud; the reduction of internal IT resource use; the organisations not dealing with maintenance activities, etc.) ● Capital expenditure reduction: due to the pay-as-you go approach to ICT with comparatively low initial expenditure and operating expenses depending on usage. ● Savings on equipment: Cloud computing replaces the need for in-house equipment, providing the advantages of data centre consolidation and shared public and private resources ● 'Utility' pay structure: users only pay for the resources they use, pay-as-you-go style via a subscription-based model ● Facilities consolidation: data centres consolidation and pooling of resources ● Better asset utilisation: due to shared resources. Organisations do not have to build for peak usage and rely solely on their own resources. Common IT infrastructure for better IT resource availability and easier collaboration ● Labour optimization: labour redirection from routine operational and maintenance duties to core competencies and mission-critical tasks ● Measured services: automatic control and optimisation of resources by metering services ● Various payment models: pay-for-use, subscription, fixed plans
3. Increased efficiency, better services and collaboration:	● Backups and data security: if a server on the network stops functioning properly, it will not lead to a complete crash, as with traditional solutions due to networked backups ● Robust resilience: automated recovery can be easier to implement in the Cloud due to more consolidated resources.

Category	Value/Benefits/Advantages
	• Speed and productivity: Cloud functionalities and tools (e.g. dashboards, real-time statistics and active analytics) can lead to faster processes and administrative burden reductions • Virtualization: the Cloud allows even small public sector organisations to innovate in ways that would not be possible were they forced to rely entirely on their own resources, facilitates improved agility & additional enhancement options • Mobility: employees can work from any location as long as they have internet, thus allowing public sector organisations to offer more flexible working hours and improve productivity • Broad reach: public institutions can use the Cloud to provide access to a wide range of services from a single online community portal at any time of day, so that stakeholders do not have to visit an office in person • Increased collaboration: real-time multi-stakeholder collaboration and co-creation from any part of the world, facilitating the creation of knowledge networks, the sharing of information and know-how across groups of stakeholders, stimulating innovation and increasing efficiency and visibility • Greater information sharing, especially during times of crisis like the COVID-19 pandemic • Application proliferation: the Cloud supports applications that offer new ways for the public sector to engage with citizens • Closer interaction: the Cloud fosters interagency cooperation and multi-stakeholder collaboration (including the public and private sector, the academia, entrepreneurial ecosystem, the civil society, etc.) on a local, national, and international level. • More efficient technical infrastructure and better data and process integration
4. Strategic value: achieving competitive advantage through cutting-edge innovative technologies and facilitation of innovation, digitalisation and business transformation.	• Streamlined work and competitive edge: Cloud providers manage the infrastructure, enabling users to focus on value-added priorities and gain competitive edge due to financial, time and other resource savings • Regular updates: to provide the most up-to-date technology • Enhanced information and service management initiatives • Business growth through innovation: cloud computing enables companies to explore new ways of creating added value • Minimising the impact on the environment: server capacity scales up and down to fit business needs, so organisations only use the energy they need reducing carbon footprints

Category	Value/Benefits/Advantages
	• In education: access to instructional courses, learning management systems and storage, virtualized collaborative technologies, etc. • In healthcare: cloud computing to connect medical devices, store medical data; collect data, analyse trends, create alerts for physicians; medicine-at-a-distance applications; etc. In government: development of govtech and smart city solutions; collection of data and analysis of trends; better document and process management, etc.

TABLE 2: BENEFITS AND ADVANTAGES OF THE CLOUD FOR THE PUBLIC SECTOR

Disadvantages. There are some disadvantages and constraints of the Cloud that could affect its effective and efficient use and implementation by the public sector, such as:

- Dependency on the provider and vendor lock-in: e.g. in regards to security issues, legal agreements, customer support, speed and quality of services, etc. Vendor lock-in is one of the biggest disadvantages of the Cloud. Organisations may face problems when transferring their services from one vendor to another. As different vendors provide different platforms, that can cause difficulty moving from one Cloud to another
- Dependency on Internet connectivity
- Cost variations: potential pricing changes, needs to purchase costly features, etc.
- Speed: dependency on the internet connection and the number of users accessing it simultaneously
- Disruption of services: due to issues at the provider's end, e.g. power outages, lack of Internet, server problems, cyber-attacks, the provider going out of business, etc.
- Security concerns regarding the level of security guaranteed by the provider; potential hacker attacks on Cloud vendors; data confidentiality, integrity and deletion, etc. Given the potential risks, the legal framework demands on Cloud providers is usually high.

Ultimately, despite these drawbacks, the cloud provides numerous advantages that outweigh the disadvantages and is definitely a disruptive technology that is here to stay and facilitate the global economy and digital transformation. Cloud technologies are constantly evolving to further support the public sector (and beyond) in meeting the challenges of the 21st century, and it is up to various organisations to make optimal use of the opportunities provided. To do this, they seek to implement various cloud-based solutions, first considering the capabilities and limitations of the cloud, having a strategic, analytical approach to their implementation in line with their set goals and priorities, and by using virtualization to support the scalability and flexibility of resources offered by cloud computing [10].

2.1.2 Definition, Value and Benefits of Big Data in the Public Sector

2.1.2.1 Big Data: Definition and Essential Characteristics

Big Data. The public sector, the private sector, the business and entrepreneurial ecosystem, NGOs, academia and the civil society constantly generate large and complex data from vast amounts of different sources, which is hard or impossible to process by standard methods, known as 'Big Data'. It can also be defined as high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making [15]. Big Data can be user-generated, personal, structured and unstructured. In the early 2000-s, the analyst Doug Laney defined Big Data as the three V-s [16], [17]:

- Volume: data is being generated by a great number of sources and institutions (e.g. sensors, smart devices, business transactions, public sector, social media, etc.); it can be captured, memorised, and accessed in huge quantities
- Velocity: with the growth in innovation, tech and IoT, data streams at an unprecedented speed and must be handled in a timely, effective manner in near-real time
- Variety: structured and unstructured data comes from various sources and in various formats

There are other dimensions of Big Data as well, such as [17]:

- Variability: data flows are unpredictable – changing often and varying greatly. It's key that organisations could predict and manage data loads (on a daily, seasonal, etc. basis)
- Veracity: the accuracy and quality of Big Data – as it comes from various sources, it's challenging to link, match, cleanse and transform data across systems
- Speed: data should and can be analysed at high speed in real time, as it becomes obsolete fast
- Value: how much Big Data is worth as an asset

Open Data. Data is open if anyone can use it, reuse it, and redistribute it, subject only to the condition that any further reuse or redistribution must take place on the same terms, and that the source may have to be acknowledged [18].

2.1.2.2 Value and Benefits of Big Data for the Public Sector

Big Data can be considered the 'new oil' of the XXI century for the public sector (and not only), and a key factor for data- and knowledge-based policy-making. More and more public organisations focus on developing data-centric and data-based policies to facilitate and promote green and sustainable growth and development, innovation and digitalisation, process transparency, good governance, citizen involvement in public decision-making, etc. Overall, the benefits of Big Data in the public sector can be grouped into **three major categories,** based on a classification of the types of benefits [19]:

- Big Data analytics: use of advanced analytics via automated algorithms for the analysis of various datasets to support data-driven problem-solving and decision-making, to reveal patterns and to make forecasts (e.g. for resource and process optimisation, risk minimisation, more efficient and inclusive policy-making, etc.)

- Increased effectiveness: use of Big Data to facilitate process effectiveness and internal transparency, allowing the public and private sectors, the academia, NGOs, businesses and residents to make more informed decisions and to develop new products and services, as well as to support the overall socio-economic situation and the quality of life
- Increased efficiency: use of data to achieve improvements in efficiency; provide better services and continuous improvement based on the personalization of services and learnings from the performance of such services. This will allow for the use of less resources (time, money, effort, etc.) to provide better services

Some concrete examples of potential benefits for the public sector may include:

- Socio-economic analysis: use of data for more efficient economic and financial forecasts
- Policy-making: more effective and efficient data-based decision-making in various spheres
- Cyber security: collection and analysis of crucial sensitive data to facilitate the detection, countering and prevention of potential cyber-attacks, and cybersecurity development
- Security and radicalisation: use of data to facilitate the detection, countering and prevention of security threats; design of more effective and efficient policies to address radicalisation
- Supporting innovation: encouraging digitalisation, innovation, and the development of new solutions via Big and open data
- Open government and data sharing (in line with strategic priorities): information sharing for better transparency, visibility and coordination, building trust between the public sector, the other sectors and the citizens
- Better and more accessible services: tailored and digitised secure services (in line with GDPR and other privacy requirements) to increase usability, effectiveness, efficiency, and citizen satisfaction (e.g. in e-services, education, social services, environment, health, security, etc.)
- Improved urban policies: collection and analysis of data (incl. generated by smart city and IoT solutions) for better urban policy-making as a key factor in improving the overall urban environment and quality of life (e.g. data on air pollution, waste, traffic, etc.)
- Citizen sentiment analysis: analysis of data on citizen sentiments from crowdsourced datasets and the media for improved priority and goal setting, awareness of citizens' needs, etc.

Constraints. Of course, there are some constraints that could affect the effective and efficient use of Big Data sources, capabilities and technologies by the public sector, among which:

- Lack of vision/awareness/willingness on a senior political level to use Big and open data
- Complex public administration bodies with lengthy procedures preventing the timely creation and realisation of a common vision and strategy related to Big Data
- Lack of skilled Big Data and business experts interested in contributing to Big Data initiatives in dealing with public sector challenges

- Complex administrative and regulatory procedures and uncertainty regarding future Big Data regulations and requirements, including on data security and protection, confidentiality, anonymisation, etc.
- Lack of communication and collaboration between various stakeholders from the public, private, academic, NGO, public sector, including exchange of experiences, know-how, expertise, etc.

However, more and more public organisations realise the importance and the value of Big Data, generated and collected by them via various daily activities (e.g. tax collection data, traffic data, health data, utilities data, air quality data, etc.), and put effort into making it more accessible and usable by different stakeholders on their own speed, thus increasing visibility and encouraging local ecosystems to develop innovative products and solutions and to contribute to the utilisation of this data. In order to use the benefits of Big Data, a key factor is that the public sector has a clear view of its long- and short-term goals and preferably a data strategy defining those, also identifying the main data sources, the processes related to the access, management and storage of data (the effective and efficient management of data through its whole life cycle also among the important challenges facing the public sector and all other organisations generating Big Data); the ways in which the data will be analysed and which policies and decision-making processes will it support, etc. The public sector can support the development of a data ecosystem via sharing of data, opening datasets via open data platforms, and by managing, facilitating and encouraging the use and dissemination of data. The public sector can also collaborate with different stakeholder groups on various open data initiatives, including the creation of open data policies, strategic documents, platforms, campaigns to foster awareness of Big Data and Big Data technologies and boost their spread and use.

In relation to what has been said so far, we should point out the **achievements** in this area – and one of the best innovations are data spaces and data space ecosystems, which are being often used in the public sector in Europe. Their focus is on the overall use of data, the life cycle of individual databases and their reuse. The benefit from the reuse of already available data, which carries information and possibilities for use even at a later stage, has been proven. The life cycle of the data is very important and complex, as well as the benefits of using the data repeatedly. In the process of working, various organisations come to the conclusion that when they combine the data they have, they can increase their efficiency. In this regard, a point is reached when different organisations come together to provide these data and services under one management, the concept of Data Space emerges. Data spaces are a data management abstraction that aims to solve some of the problems with data integration systems. The topic of data spaces is covered in detail in *Section 3.6 Data Space Ecosystems in the Public Sector*.

2.1.3 Cloud and Big Data: The Value and Benefits of PolicyCLOUD

PolicyCLOUD aims to harness the potential of digitisation, Big Data and Cloud technologies to improve the modelling, creation and implementation of public and business policies by delivering a unique, integrated environment of curated datasets, and data management, manipulation, and analysis tools. The project addresses the full lifecycle of policy management using the data analysis capabilities of the European Cloud Initiative. The solutions and tools developed by PolicyCLOUD will eventually become available as public Cloud services from the European Open Science Cloud (EOSC) [20]. By 2023, PolicyCLOUD aims to roll out a series of novel data-driven policy management solutions to drastically advance policy-making and benefit a wide variety of stakeholders, i.e. public administrations, policy-makers and NGOs, academic institutions, Cloud and Big Data solution providers, and last but not least, citizens. All of the benefits and values of the Cloud and Big Data, identified in Sections 2.1.1 and 2.1.2, can be applied to PolicyCLOUD. The project will provide 6 services to facilitate the transformation of raw data into valuable and actionable knowledge for efficient and effective policy creation [21]:

Overview of PolicyCLOUD Services	
Cloud Capabilities & Data Collection	Adaptable Cloud gateways and APIs to obtain data from heterogeneous sources, e.g. online platforms and the Internet of Things: • A Data Governance Model is integrated to ensure access to data across the complete data lifecycle • Incentives Management is created and it's an opportunity for citizen involvement
Reusable Models & Analytical Tools	A set of technologies spanning the complete data lifecycle enables opinion mining, and sentiment, social dynamics, and behavioural data analysis. Semantic annotations will be integrated to ensure data interoperability and reusability.
Policy Development Toolkit (PDT)	Automated tools enabling non-experts to manipulate, model, and visualise data inputs and the outputs from analytics processes. Visualizations cover a wide range of aspects from data analytics to policy evaluation outcomes.
Ethics Framework	A documented approach ensuring the provision of privacy and security for sensitive data.
Data Marketplace (DM)	A repository of exploitable data and knowledge to be used across the full range of policy-making scenarios.
Policies Management Framework	Decision tool based on the PolicyCLOUD "Data-driven Policy Lifecycle Methodology", enabling the integration of data collection, modelling and simulation technologies and the assessment of policy compliance and impact.

TABLE 3: THE POLICYCLOUD SERVICES

In the table below, a list of the benefits and advantages of the PolicyCLOUD project for the different groups of stakeholders from the public sector (and not only) is being introduced [22]:

Stakeholders	PolicyCLOUD: Benefits and Advantages
Policy Makers at EU, national and regional level, PAs, NGOs and Standardisation Bodies	Provision them with access to tools for efficient and effective policy decisions: • Cleaned, refined, structured and trustworthy datasets emerging from the pilot use cases • Analytical tools to enhance the predictive power of data • Scenario simulations to model and evaluate policy impacts
Academia, Research Centres, Researchers and Big Data Experts	Facilitating better quality research outcomes through access to: • Solutions and policy making services available via the EOSC • Previous project results upon which to build further

Industry (e.g. Big Data and Cloud solutions providers)	Facilitating improved efficiencies and new opportunities via access to: ● Novel data management and analysis solutions ● Tools for cleaning and refining data ● Improvements in the quality of governmental datasets via pilots ● The Data Marketplace as a shop window to offer new datasets
Citizens: residents of the pilot use cases and cities/countries impacted by future PolicyCLOUD adoption	Facilitating innovation and improved quality of life via: ● Opportunities to participate in policy creation through the iterative design and implementation methodology adopted and co-creation activities ● Continuous improvement in policies through an iterative design approach which leverages newly available data on an ongoing basis and optimises policies simultaneously across multiple sectors ● Increasingly relevant and useful policies designed around homogeneous segments rather than entire populations

TABLE 4: THE BENEFITS OF POLICYCLOUD

2.2 The Cloud and Big Data in the Public Sector

2.2.1 The Cloud in the Public Sector: Use Cases and Good Practices

This section will showcase good practices, success stories and benchmarks related to:

- the use of the Cloud in the public sector in the four PolicyCLOUD use case countries: Italy (Lombardy Region), Spain (The Autonomous Community of Aragon and Barcelona), Bulgaria (Sofia) and the UK (Camden, London), Germany (Berlin);
- the use of Big Data in the public sector in the four PolicyCLOUD use case countries: Italy, Spain, Bulgaria and the UK, as well as several international examples from the PolicyCLOUD use cases domains (e.g. radicalisation, urban environment, agriculture, employment, etc.).

The research below aims to raise awareness of and serve as an inspiration regarding the capabilities and the value of the Cloud and Big Data in the public sector, as well as to provide the first step and the basis for the development of a set of adaptable approaches for the examination of public authorities' policy-making structures and their systems maturity to absorb and analyse data in common Cloud infrastructures to be added to the methodology suggested in Section 4 in the second iteration of D5.1 in M32. The suggested analysis includes examples of political will and commitment; public institutions in charge of digitalisation and innovation on national, regional and municipal level; strategic documents and concrete projects and initiatives related to Cloud and Big Data implementation in the public sector, as they are key aspects signifying the PA's readiness and maturity to implement Cloud and Big Data solutions.

2.2.1.1 ITALY

Italy ranks 22nd in the EU in Integration of digital technology according to the 2020 Digital Economy and Society Index (DESI). The share of enterprises using social media rose to 22% (close to the EU's average of 25%). The use of Cloud services remained stable (used by 15% of Italian enterprises) and just below the EU average (18%). The use of Big Data also remained stable at 7%, close to the EU's average (12%) [23] [24]. In recent years, Italy has adopted several strategies to facilitate digitalisation, innovation, and ICT development, with special focus on the implementation of Cloud and Big Data solutions [25].

Italia 2025. The 5-year National Innovation Plan '*Italia 2025*' adopted in 2019, gives the public administration (PA) a central role as an enabler and driver of innovation and digitisation in Italy at the centre of a 'process for the structural and radical transformation of the country' [25].

The 2019-2021 Plan for Information Technology in the Public Administration, updated to a 2020-22 version, sets out a comprehensive list of actions with particular emphasis on the adoption of a Cloud paradigm in PAs to rationalise the current digital infrastructure and services and improve their security, efficiency, and reliability. The plan stresses the ***Cloud first principle***, favouring the priority adoption of Cloud technologies and solutions by PAs when developing new services or acquiring software [25], [28].

The Cloud Strategy of the Public Administration (CSPA) [25]. The Agency for Digital Italy (AgID) – the technical agency of the Presidency of the Council of Ministers of Italy [26], developed CSPA to facilitate and encourage the adoption of the Cloud model, in line with Italy's Guidelines on the Digital Growth Strategy and the 2019-2021 Plan for IT in the Public Administration [27]. The adoption of Cloud infrastructure is expected to facilitate higher operational efficiency, cost reductions, easier and cheaper software updates, improved security and data protection, faster services for citizens and businesses. CSPA aims to provide a list of prerequisites for public and private entities that want to provide Cloud services and infrastructure to the public administration in line with specific security and reliability parameters, and principles suitable for the needs of public administration [28]:

- improvement of service levels, accessibility, usability and security
- interoperability of services within the Cloud model of public administration bodies
- reduction of the risk of vendor lock-in with the service providers
- requalification of the offer, expansion and diversification of the suppliers' market
- resilience, scalability, reversibility and data protection
- opening up of the market to small and medium-sized enterprises (SMEs)

ICS also outlines the so-called ***Cloud of the PA model*** [29] with compliance requirements for security, performance, scalability, interoperability, portability, legislative aspects in regard to the technologies the public sector is looking for: SaaS, IaaS, PaaS services from different categories of Cloud providers. Providers who meet the quality requirements are included in the ***Cloud Services Catalogue of Qualified Services for the PA Bodies (Cloud marketplace AgID[1])***, where all PAs can research the available

[1] The key mission of AgID is to manage the implementation of the Italian Digital Agenda in line with the EU Digital Agenda. For more information, please check the AgID official website: https://www.agid.gov.it/

infrastructures and services and since 2019 can acquire infrastructures and solutions [25]. Furthermore, the Department for Digital Transformation, in collaboration with the AgID has developed a ***Cloud Enablement Program*** that defines the set of activities and resources useful to administrations for the migration of digital services and infrastructures to the PA Cloud[2].

According to a questionnaire aimed at showing whether Cloud solutions are a priority choice for PAs as part of the **Report to Parliament on the State of Implementation of the 2017-2019 PITPA** by the Italian Supreme Audit Institution (2020), 51% of the participating 7,237 entities reported that they use Cloud services; this number increased up to 91% in the regions/autonomous provinces. Among the local authorities, the Municipalities with less than 5,000 inhabitants declared values below 50% [30].

2.2.1.2 Spain

Spain ranks 13th in the EU in Integration of digital technology according to the 2020 DESI Index, which shows that Spanish businesses take advantage of the opportunities presented by digital technologies in line with the EU average: 43% of enterprises have an e-information sharing system (against 34% in EU) and almost 1/3 use social media; 16% of companies use the Cloud (18% in the EU) and 11% - Big Data analysis (12% in EU). Initiatives related to digitisation, AI and other emerging/disruptive technologies can further boost the innovation capacity of Spain, which is committed to advancing innovation and investing strategically in digital technologies, including via EU programmes [31].

The Digital Transformation Plan for the General Administration and its Public Agencies (also known as ***ICT Strategy 2015-20***) defines the ICT Strategy and the main activities to be followed by all the ministries in Spain. According to the Strategy, all digital services of the central state administration must have a common e-Identification administrative system ***(eIDAS)***, accessed via shared keys [32]:

- **Cl@ve** is the common platform for identification, authentication and e-signature, a horizontal and interoperable system linked to the national eIDAS, complementing the existing e-systems for public services access via electronic ID cards and e-certificates. It offers e-identification via shared keys and signing in the Cloud with personal certificates kept in remote servers. Approved in 2014, Cl@ve was adopted by 7000+ public organisations and reached 8.5+ mln. registered users and 172 mln. operations by 2020. Currently, the Spanish government is incorporating Cl@ve in all digital services that need e-Identification for all PAs and can be used by any PA from any administrative level.
- **@firma** is a MultiPKI national validation platform providing free eID and e-signature services to eGovernment applications, allowing the verification of the state and validity of all qualified certificates allowed in Spain, and the e-signatures created by citizens and businesses in any eGovernment service. **@firma** is offered as a Cloud service to various eGovernment services as open software to be installed by PAs with a high demand of signature services, and can also be accessed via **Cl@ve**. 863 million+ transactions were made through **@firma** in 2019.

[2] Additional information about Italy’s Cloud strategy can be found on the dedicated website Cloud Italia: the main source for the development of Cloud computing in the digital infrastructures of the Italian public sector https://cloud.italia.it/

A special **National Plan to Stimulate the Demand for Cloud Solutions** has been developed by Red.es to encourage the adoption and expansion of the Cloud in Spain. Red.es is a public corporate entity belonging to the Ministry of Energy, Tourism and the Digital Agenda, and plays a key role in executing the 2025 Digital Agenda for Spain, focused on the digital convergence with Europe, the optimisation of public services and the development of the digital economy [33]. The Plan aims to support SMEs by providing financial assistance to firms in the ICT sector with turnover below €50 million for 9 months in order for them to adopt Cloud solutions [32].

2.2.1.2.1. The Autonomous Community of Aragon

Currently, the Autonomous Community of Aragon is working on its own **Bill of Measures for the Implementation and Development of Cloud Computing (BMIDCC).** The promotion of Cloud computing with special emphasis on its application for the development of new solutions as well as the promotion of its use in various industries (e.g. healthcare, education, governance, research, business, etc.) is a key goal, set in the **2020 Strategy for Social and Economic Recovery of Aragon**, proposing measures to deal with and recover from the consequences of COVID-19. Furthermore, one of the global Cloud leaders is expected to build 3 data processing centres and associated facilities in three Aragon municipalities, which also requires the availability of certain Cloud-related regulations.

As of now, no regulatory framework for the Cloud exists in Aragon, and the Bill aims to form a strategy, propose a series of guidelines and establish measures that will support the implementation and development of cloud computing in Aragon, including the creation and application of a Cloud Policy by the Aragonese Government, as well as Cloud-related standards and regulations. The initial public consultations for the preparation of a draft bill were completed in the spring of 2021. By developing BMIDCC, the PAs of Aragon aim to promote and facilitate the development of digitalisation, innovation and technology, and encourage the use and implementation of the Cloud as a key technology to boost the local ICT and innovative sector and to support the region's productivity and competitiveness [34].

Cloud-based recommendations system. The Aragon Regional Government (ARG) developed a unified e-services portal where citizens can apply for 1500+ administrative procedures. It also includes a recommendation system, aimed to improve citizens' interaction with public services, based on semantic processing of information from similar or related electronic procedures, institutional websites managed by the ARG and ARG's Twitter accounts. The information from Twitter is collected using a special SaaS-based social media data mining application integrated with MiCADO framework to mitigate the need for an oversized infrastructure to cope with potential unpredictable fluctuations of CPU load. The recommendation system is planned to be extended to include additional sources of information, e.g. the historical behaviour of applicants [35].

Virtual desktops. In May 2020, after the beginning of the COVID-19 pandemic, the public agency Aragonese Telematic Services (AST) together with IECISA Consulting, provided the agriculture, employment and the general administrations of Aragon with 1000 customized virtual desktops using Amazon WorkSpaces and a pay-as-you-do model to ensure the continuous remote work during the pandemic. For less than 2 weeks, the project was integrated with AST's directory based on free software.

The key aspects considered were security and access to AST's own applications and corporate directories in the shared storage; scalability, simplicity, availability and support. A virtual private cloud was created to host WorkSpaces and was connected via VPN to the corporate environment, from where AST provides services to ARG. Thus, WorkSpaces could access shared application data and allow officials to use all their corporate applications, directory and shared storage the same way that they would if they were in their offices. The virtual desktops do not require prior capacity analysis, additional components, unwanted OS, creation of networks, VPN, or additional security, and is as simple as sending a link to connect both from a mobile device, a tablet, or a traditional desktop [36].

2.2.1.2.2. Barcelona

Cloud spaces are so widespread in Barcelona that their application is reflected in every area of public life. Cloud Cities Barcelona was created in May by Tomás Saraceno, inviting participants into a space of communal encounter, discourse and critical speculation, suspended at a height of over 130 metres above the city of Barcelona.

The administration of Barcelona is making efforts to encourage young people to become more educated in the field of digital opportunities, to acquire new skills that they can put into their work.

Microsoft Cloud services and devices play a transformative role in supporting several of the city's initiatives, from providing data centre management solutions that increase city employee productivity to hosting and analysing public data. Activities that are unthinkable without cloud space. Through new technologies, the City Council is reinventing Barcelona, using new technologies as pillars of future growth and sustainability.

Chosen by the GSM Association (GSMA) as the Mobile Capital of the World until 2018, Barcelona expects to become one of the leading "smart cities" in the world. Through its work with the City Protocol Society and technology leaders from industry and academia, Barcelona is helping to promote new standards and innovations for smart city services that are being adopted by other cities around the world.

By adopting the latest identity, security and device management tools, Barcelona City Council improves services for citizens by providing city employees with the latest technology they need to do their jobs.

Most city employees use standard desktop or laptop computers, but the city has also implemented a remote desktop computing model (in this case, virtual desktop infrastructure, or VDI) to better meet the needs of on-site workers. Users can access their desktops from a desktop, laptop, thin client (devices with less powerful processors and less memory), tablet or smartphone. The benefits of desktop virtualization include improved centralised control over desktop environments, ubiquitous access to applications and data, and lower desktop and client management costs.

2.2.1.3 Bulgaria

According to the 2020 DESI Index, 23% of Bulgarian enterprises have an electronic information sharing system in place (against the EU average of 34%), 7% of companies use the Cloud (18% in the EU) and 6% access Big Data analysis (vs EU's average of 12%, and close to Italy's 7%) [37]. The public sector in Bulgaria recognizes the importance of the Cloud and Big Data for e-government, economy and quality of life, and is focused on developing appropriate conditions for their implementation and use.

In 2022, with the amendments and supplements to the Electronic Government Act adopted by the National Assembly, the State Agency for Electronic Government was closed and the Ministry of Electronic Government was established. By Decree № 12 of February 4, 2022 of the Council of Ministers, the Rules of Procedure of the Ministry of Electronic Government were adopted. SG no. 8 of 2022.

Architecture of the E-government in the Republic of Bulgarian. The architecture of the E-government is an integral part of the implementation of the E-government policy [38] defined in the Strategy for Development of e-Government in the Republic of Bulgaria and the Law on E-Government. The implementation of the document is subject to effective coordination between all stakeholders. Approved by the Chairman of the State Agency for Electronic Government with Order № DAEU-5040-11.04.2019, the Architecture of Electronic Government in the Republic of Bulgaria is aimed at the implementation of:

1. digital transformation of the administration
2. mandatory use by the administrative bodies of the horizontal systems and the shared resources of the e-government;
3. mechanisms for coordination and control of the implementation of the architecture;
4. application of uniform standards and interoperability in the design, construction, upgrading and implementation of information solutions;
5. defines the participants in e-government, their functions, the principles of EU and the requirements for system and technological architecture.
6. sustainable high overall level of network and information security;
7. transformation of data into information and knowledge;
8. gaining trust from citizens and businesses.

Updated Strategy for Development of e-Government in the Republic of Bulgaria 2019 – 2025 [40]. The Strategy for the development of #-government in the Republic of Bulgaria covers the development of four main areas of communication and services:

1. "Administration - Citizens" - includes modern Internet and intranet WEB-based solutions, combined with traditional means of providing wide access, leading to qualitative changes in the conditions for communication and provision of services to citizens.
2. "Administration - Business" - includes modern solutions that optimize the processes and business relations between the administration and various economic entities.
3. "Administration - Administration" - includes the development of information technology in national and interstate aspects in order to effectively interact between different administrative structures.
4. "Internal efficiency and effectiveness" - includes the organization and optimization of business processes, the relationship "Administration - Employees" and communication in the various administrative structures.

Updated Roadmap for the implementation of the Strategy for the development of e-government in the Republic of Bulgaria for the period 2019 – 2023. The Roadmap is a document that includes the measures and activities to implement the strategic objectives set out in the Updated Strategy for Development of e-Government in the Republic of Bulgaria for the period 2019 - 2023, as well as the responsible institutions and the necessary financial resources.

Strategy for the Development of E-government in Bulgaria 2020-25 (SDEBG). DAEU was in charge of the development of the SDEBG 2014-19 and the updated 2020-25 version [40], setting out the general framework for the development of e-government in Bulgaria, including the implementation of SPHC and Big Data solutions, a unified portal for e-services, open data platform, cybersecurity solutions, etc.

State Hybrid Private Cloud (SHPC). The achievement of a fully functional shared SHPC with fully migrated administrative information systems of administrative bodies and public institutions by 2025 is among the ambitious priorities set in SDEBG. The development of SHPC and the secure Internet connections it needs are key for building a highly sustainable Cloud environment, located in two data centres: an e-Government Control and Technical Centre and a Data Recovery Centre providing IaaS. The SHPC is being currently updated and its capacity gradually expanded, as it will allow for significant economies of scale; ICT resource consolidation and centralisation; centralised maintenance of shared resources and cost optimisation; secure, flexible way to provide Cloud resources and services (e.g. for in-house information and communication, projects and systems for information arrays and databases of national significance; IaaS, SaaS, PaaS, and Government-as-a-Service/Platform services, etc.) [3] [40].

Cloud Electronic Signature (CES). The Bulgarian E-Identification Act envisions the development of a national e-Identification Scheme, which is currently being developed as part of the ***Unified Model (and Portal) for Access to Electronic Administrative Services of DAEU (UPDAEU)***. In 2019, in support of this Scheme, DAEU introduced the CES with a two-factor identification system for more information security in addition to the already existing e-ID tools. CES allows citizens and businesses to request administrative services and to access them 24/7 from any point of the world via the internet.

MGovernment. In line with the EU priorities, Bulgaria is making steps towards MGovernment – the provision of e-services fully accessible through mobile platforms and devices. MGovernment includes the creation of an infrastructure for managing the identity of mobile devices without imposing restrictions on users and the availability of services via chatbots, AI and Cloud services [40].

Strategy for Using Cloud Solutions in National Security 2014-20. Developed by the Ministry of Transport, Information Technology and Communications (MTITC), it aims to support the efforts on national level to

[3] SDEBG defines SHPC as a centralized state information infrastructure (servers, data storage facilities, communication equipment, engineering and technical equipment and system software), distributed in several locations in premises meeting the criteria for construction of secure information centres, which provides physical and virtual resources for use and administration by PAs, while ensuring a high level of security, reliability, isolation of individual users and the impossibility of interference with the performance of their information systems or unauthorized access to their information. SHPC will comply with the key principles of digital sovereignty, security, interoperability, reusability and with all EU legal and ethical provisions; it will use common standards, modular architectures and open source and make software, data, tools generated by the PA publicly available

improve PAs via Cloud information technologies and the transition from the provision of e-communications to integrated e-services, and in particular in regard to national security.

Digital Bulgaria 2025 National Programme. Developed by the MTITC, 'Digital Bulgaria 2025' aims to facilitate digitalisation and innovation in all socio-economic areas by facilitating an innovative environment and unified standards for better security/interoperability. It sets strategic priorities, e.g.:

- creation of conditions for the development of digital networks and services and better access to Cloud, Big Data and IoT services, incl. via the development of a strong, competitive and dynamic telecom sector to exploit the potential of innovative technologies
- development of a dynamic and innovative digital economy and increasing its competitiveness via investments in ICT, cloud computing, Big Data, etc., and support of research and innovation
- digitization of the industry and development of a data-based economy by encouraging the use of innovative tech (e.g. Cloud, blockchain, AI, Big Data) and by developing policies [41].

2.2.1.3.1. Sofia

Innovative Sofia. The Digitalization, Innovation and Economic Development department of the Sofia Municipality (Innovative Sofia), headed by a Deputy Mayor, was established in 2020 to support the development of Sofia as an intelligent, digital, innovative and tech city. The goal of Innovative Sofia is to be a leader and engine of the digital transformation of Sofia: to initiate and coordinate digital transformation policies; to consolidate and guide the city's digital and smart city projects; to facilitate the use of Big Data and Cloud solutions and the development of e-services; to support the high-tech and R&D sector, international partnerships and a recognizable international brand of Sofia [42].

Digital Transformation Strategy for Sofia (DTSS). Sofia Municipality developed DTSS and the Action Plan for its Implementation (APDTSS) together with 100+ representatives of the local ecosystem and EU experts. SDTS aims to encourage the development of innovative solutions for digital transformation, and to support the ecosystem enabling the development of new digital and smart solutions. It recognises the importance of a public repository of solutions, a government cloud (G-cloud) for the development of new solutions and recommends the use of SOA for the construction of software for distributed computing, e-services and integration of applications. APDTSS sets the key projects that Innovative Sofia should realise by 2022, and the list is being periodically expanded and updated [43].

Cloud E-services for the Administration. In 2021, Sofia joined the project together with DAEU, and three other cities. The 27-month project will ensure the implementation of an SPHC-hosted platform with tools to improve the administrative workflow and shared ICT services. The goal is to provide citizens with access to more e-services delivered faster via unification and integration of municipalities and the central administration. A Catalogue of 'E-administration' services with the individual work processes to be automated through the platform will also be defined, supplemented and accessible through the automation platform, and personalised packages will be configured initially for the three pilot municipalities. After that, other administrations in the country will also be able to join [44].

School in the Cloud. Part of the Sandbox for Innovative Solutions of Innovative Sofia, this pilot project aims to support digital skills development and work in an online environment, as well as the integration of the 'One-to-One' model for managing the learning process via a device for each teacher and student and special trainings in order to facilitate the overall digital transformation of schools. So far, over 200 teachers from 16 schools have joined the project [45].

2.2.1.4 UK

According to the 2020 DESI, the UK ranks 8th in terms of integration of digital technology with 19% of enterprises having an e-information sharing system (against the EU average of 34%), 30% of companies use the Cloud (18% in the EU) and the latest data on access to Big Data analysis is from 2018 – 15% [46].

Cloud Standards. In terms of Cloud standards, the UK follows the guidelines from the UK's National Cyber Security Centre (NCSC) and the international standard for information security ISO27001. As ISO27001 has not been formally adopted in the UK, it's used only for guidance.

- NCSC is an organisation of the UK Government providing advice and support for the wider public and private sector, as well as the general public, combining the efforts of academia and industry to improve cyber security. By providing information in an accessible way, the NCSC seeks to inform citizens and organisations about the opportunities available in today's digital world, how to avoid computer security threats, and in case of incidents provides effective response to minimise harm, help with recovery, and learn lessons for the future [47].

Cloud Security Guidance. The NCSC has published a Cloud Security Guidance aimed at the public sector and enterprise organisations showcasing how to configure, deploy and use Cloud services securely. The Cloud Security Guidance proposes accessible information related to Cloud security and builds a framework to help users evaluate the security of any Cloud service, built around 14 main principles and 8 key steps. The Guidance helps Cloud users determine whether a Cloud service is secure enough to handle their data and gives Cloud service providers recommendations on how to present the security properties of their offerings to the public sector and enterprise clients [48], [49].

Government Cloud First Policy. The UK government introduced a 'Cloud First' policy in 2013 for all technology decisions. This approach is mandatory for central government and strongly recommended to the wider public sector, with preferably focusing on a public cloud rather than a community, hybrid or private one. The idea behind the policy is that when procuring new or existing services, public sector organisations should consider and fully evaluate potential Cloud solutions first before considering any other options. Departments could choose an alternative solution, but first need to prove that it offers better flexibility, security or value for money [50].

GOV.UK Platform as a Service (GOV.UK PaaS) is a security-compliant Cloud-hosting platform built by the Government Digital Service (GDS) for running public sector services in the Cloud, which offers tools to help teams build and run Cloud native applications in production. GOV.UK PaaS is a life service, managing the deployment of apps, services and background tasks, hosted in London and Ireland. It uses the open-source Cloud Foundry project, and runs on Amazon Web Services, with applications being automatically run across three independent availability zones to ensure resilience [51].

Digital Marketplace (DM). DM is an online platform and the single place that all public organisations in the UK can access to find and buy different Cloud-based services, reach out to specialists who can work on digital projects and provide expertise in various fields, as well as physical data centre space for sensitive data requiring high security and legacy systems that cannot be migrated to a Cloud.

The **Digital Marketplace Strategy** behind the DM and the DM itself aim to transform the way the public sector buys digital and technology services, making it simpler, clearer and faster to buy what it needs, and providing suppliers a transparent and fair way to offer their services to the UK public sector, facilitating jobs creation and economic growth. Furthermore, the DM has allowed SMEs to gain better access to the public sector and has helped drive supplier competition [52].

The UK Government Cloud Strategy (G-Cloud). Published in 2011, it is as a sub strategy to the Government ICT Strategy, aimed to transform the government ICT landscape and create a more productive, flexible workforce delivering better digital public services. Through the Strategy, the UK Government confirms its commitment to adopting the Cloud, so that the ICT sector could become more agile, cost effective and environmentally sustainable [53].

2.2.1.5 Germany

Germany is among the top places in the world in terms of the development of digitalization. The availability of specialists in the IT sector is ranked lower by international publications. This in itself shows the possibilities of digitization and the efforts that are being made to further develop the digital presence of the country, facilitating the daily life of citizens through digitization. It is a leader in terms of preparations for the launch of the 5G network and also boasts a fairly wide availability of broadband internet - although not the highest bandwidth.

According to the information shared by the European Commission on the results of the study of the use of cloud spaces in the public sector, it is clear that Germany has significantly increased its investments in this direction. They invest a lot of time and resources in the direction of improving the security of cloud spaces, their sustainability and application.

Digital Strategy 2025. Germany has its Digital strategy 2025. It describes the priorities of the German government, namely the development of digital capabilities and the promotion of the use of new tools in order to improve the digitization processes in Germany. The strategy is based on 10 pillars important to digitization, including a pillar that focuses on the introduction of digital education and throughout all stages of a person's life.

The main objectives of the digital education pillar are as follows:

- By 2025, every student should have basic knowledge of informatics, how algorithms work, and programming. In this regard, appropriate courses are introduced in the curricula of primary and secondary schools and in the training and continuing education of teachers.
- By 2025, Germany will be one of the leaders in digital infrastructure in the education sector.

- By 2025, the workplace should be the number one place to acquire the latest information technology (IT) knowledge.
- By 2025, all publicly funded educational institutions must provide core learning materials online. To achieve these goals, we must push for education in and for the digital world in 2025 at all levels – starting from school, including the dual system of vocational training, to university and vocational continuing education.

Another very telling factor for Germany's digital development is the availability and integration of cloud spaces. In this regard, a number of studies have been carried out on the benefits, implementation and application of cloud spaces in the daily work of companies in the country. One piece of research done on the cloud in the German economy, and according to a study by Lünendonk and Hossenfelder, only half of the IT resources needed in 2019 come from the cloud. To date, however, about two-thirds of companies surveyed are currently in the process of moving applications from their data centres to the cloud. According to the survey data provided, the cloud is now almost standard: 78% of German companies used it in 2019 – after two years of an increase of 5%, which will continue. And precisely specialised departments increasingly rely on IT support from the cloud.

2.2.1.5.1. Berlin

Leader in Digitalisation. Berlin is a good example of a city's development in digitalization. They have a lot of practices related to the field. The municipality of Berlin pioneers innovations in data processing, development and delivery of electronic systems, manufacturing and microelectronics. The underlying objective is to build on the synergies between these areas and thereby cater for the entire chain of digital network technologies. Berlin is the ideal place for this purpose because it is a centre that brings together digitization activity at different levels and can process and model the results achieved. Such projects have access to the institutes' existing technological infrastructure and offer the advantage of direct transfer of research results into practice. The Berlin Centre for Digital Transformation works closely with the Technical University of Berlin, in particular with the faculties of Mathematics and Natural Sciences, Electrical Engineering and Computer Science, and Transport and Machine Systems. The concentration of data in one place drives digital networks in Germany and abroad and positions Berlin as a centre for the next wave of digitization.

Digital Twins. The municipality of Berlin already created digital twins. They have become valuable in manufacturing and other areas are also experimenting with all information for the city, they have all date and can make different analysis, they can make radical analysis of the already collected data in order to be able to build policies that lead to positive results in the direction of improving the urban environment.

The Berlin Centre for Digital Transformation is a joint venture involving the four Berlin-based institutes Fraunhofer FOKUS, HHI, IPK and IZM. The goal here is to find solutions that recognize the growing digitization and networking in every area of our lives. The centre concentrates its activity mainly on research divided between core and interface technologies and solutions in four specific application areas:

- Health and medicine
- Mobility and the city of the future

- Industry and production
- Critical infrastructure

2.2.2 Big Data in the Public Sector: Use Cases and Good Practices

2.2.2.1 International Use Cases and Good Practices

Data.europe.eu. is the official portal for European data providing access to open data from international, EU, national, regional, local and geodata portals. It replaces the EU Open Data Portal and the European Data Portal, and currently includes over 1.4 million datasets and 170 catalogues from 36 countries in 13 sectors (e.g. agriculture; economy; education; environment; government; health; transport, etc.). The Portal also collects the metadata of public data made available across Europe [54].

OpenTraffic is a global data platform to process anonymous positions of vehicles and smartphones into real-time and historical traffic statistics, built using fully open-source software [55]. The World Bank is a founding partner of the platform alongside the non-profit OpenTraffic, coordinating open-source software development and data collection by individuals and organizations across the transport sector

Uber Movement. Uber Movement shares open, interoperable anonymised data aggregated from 10+ bln. Uber trips in 700+ cities to facilitate urban planning around the globe. It provides free downloadable data aligned with open standards (for easier integration) and tools for cities to more deeply understand and address urban transportation challenges from infrastructure planning to mobility research and to facilitate reduced congestion and emissions, and improved road safety [56].

Global Terrorism Database (GTD). GTD is the most comprehensive unclassified open-source database of terrorist attacks in the world, containing information on 200K+ terrorist attacks occurring between 1970 and 2015. It's managed by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) so that terrorism can be more readily studied and defeated. For each event, a wide range of information is available, incl. date/location, the weapons used, number of casualties, etc. [57].

2.2.2.2 Italy

In the last decade, Italy has been very active in its efforts towards open and Big Data. Strategies for open and Big Data have been published both on national and regional level, with some of the regions having their own open data rules, provisions and dedicated portals, including the Lombardy region, which is the focus of PolicyCLOUD's first use case.

Italian Digital Government Code (DGC). It's interesting to mention that while open data are defined in Italian law under the Italian DGC, Big Data has no definition in Italian law [18]. Under DGC, each regional open data law entrusts the regional government with the promotion of public data digitalisation via cooperation with regional and local governments for the creation of a uniform national open data standard. Thus, regional laws suggest measures to support the PA to carry out open data policies.

The 4th National Open Government Action Plan 2019-21 (NOGAP). Italy was among the first countries to join the international **Open Government Partnership (OGP)**, promoting the adoption of public policies based on transparency, participation, and innovation. As part of OGP, the Italian Department for Public Administration and the OGP team have developed four 3-year NOGAPs since 2010. The 4th NOGAP is a comprehensive strategy for open government setting 19 actions to encourage open data, transparency, corruption prevention, digital citizenship & skills, etc., including:

- setting measures for the adoption of a national licence for appropriate and effective data reuse and setting of common standards for types of data released,
- releasing a growing number of high-value open datasets and tools to support their exploitation, including data visualisation solutions and documented API
- developing joint open data initiatives to increase transparency and accountability and foster data reuse; enhancing the quality and quantity of released data in an open format, etc. [58]

Guidelines and policy recommendations for Big Data (GPRBD). In 2019, the Italian Competition Authority, the Authority for Communications Guarantees and the Authority for the Protection of Personal Data published GPRBD to facilitate the development of a digital economy based on the acquisition and analysis of Big Data and for privacy, regulation, antitrust and consumer protection [59].

#Data4All. An initiative by AgID to facilitate the access to specific open data of the Italian PAs available in 3 thematic portals. **#Data4All** provides a single information point about 3 open data projects of the Presidency of the Council of Ministers and allows for: better communication with citizens and the public sector; more effective and efficient use of datasets incl. sharing, using and reusing data; fighting corruption; better awareness of public spending and open data, etc. **#Data4All** was a key focus of the **Italian National Open Data Portal (INODP)** that used to filter the metadata of the 3 platforms [60]:

- **SoldiPubblici** gives access to public spending data from all local administrations to increase citizen participation in public management, improve access to data, enhance transparency, and allow PAs to rationalise their spending and compare expenses. The platform also provides an interactive map and is included in the new Italian Freedom of Information Act [61].
- **ItaliaSicura** is a portal for monitoring government actions aimed at countering risks related to natural disasters. It provides an overview of the areas under a status of national emergency and provides access to emergency-related open data and the related government activities [62].
- **OpenExpo** provides data on the management and organisation of Expo 2015 to facilitate accuracy and public spending transparency. The model could be potentially reused for other key events [63].

INODP was developed by AgID in 2011 and contains 150+ public open databases made available by PA bodies in various fields, e.g. economy and finance; education, energy, environment, public sector, healthcare, regions and cities, science, etc. [64]. The data are being updated and are available for analysis, study and development of apps. AgID continuously develops the portal so that it could include both metadata describing open data of PAs and databases owned and managed by public bodies. The metadata on the portal are provided as a JSON-LD open format compliant with DCAT-AP, for the automatic harvesting with the European Data Portal to enable a cross-data portal search for data sets and make public sector data more searchable. INODP also supports CKAN API to be used through HTTP requests and to provide the results in JSON [60].

Big Data and statistics. The Italian Institute of Statistics (ISTAT) set up an investigative panel on the effective use of Big Data in putting out official statistics, which prepared a specific roadmap. Furthermore, the Authority for the Protection of Personal Data authorised ISTAT to analyse Big Data gathered from mobile devices for the purpose of compiling national statistics [18].

Km4city. Km4city is a platform developed by the University of Florence aggregating open, private, static, and dynamic data about the Tuscany region, collected from various portals and the open data of regional municipalities. It is interoperable, enables "smart" data querying and creation of apps, and provides a unified system for the development of a "knowledge-making model" for Florence [18],[65].

2.2.2.2.1. Lombardy Region

Open Data Lombardy. The open data portal of the Lombardy region (key focus of PolicyCLOUD's Use Case 1) provides 5000+ datasets in 22 categories, among which agriculture, culture, trade, energy, government, education, mobility and transport, health, university and research, statistics, etc. [66].

2.2.2.3 Spain

Digital Spain 2025 Strategy. In 2020, the Spanish Government adopted this 5-year Strategy aimed at supporting the country's digital transformation and economic growth via better digital connectivity, improved cybersecurity, digitisation of PAs and companies, and innovation. One of its key strategic axes is the movement towards a data economy, guaranteeing security and privacy, with the aim that by 2025 at least 25% of Spanish enterprises should use Big Data and AI [67].

The 4th Action Plan of Spain on the OGP (2020-24) Spain is part of OGP, just like Italy, and also prepares action plans under the Initiative. For the first time, the 2020-24 Plan featured not only commitments from the State General Administration, but also the most innovative initiatives of the autonomous communities and cities/local entities. Some of the objectives set in the plan include:

- Facilitating transparency, open data and accountability of PAs
- Promoting data openness through the transposition of the (EU) Directive 2019/1024 on open data and the re-use of public sector information, also known as the 'Open Data Directive'
- Development of a ***Transparency and Accountability Improvement and Reinforcement plan*** as a coordinated strategy for the reinforcement of transparency and accountability and open data
- Open data strategy for gender equality policies (Catalonia)
- Creation of an Open Registry of Climate Information of Spain – (RAICES), supplying climate data from various networks operated by the Spanish PA, the business and the citizens in general
- Development of a ***Data Openness Plan***, taking advantage of the opportunities offered by Big Data. Modelling of data sets for exploitation with Big Data technologies (Andalucía).
- The commitments of the Aragon Regional Government (focus of PolicyCLOUD's second use case) include: a programme of collaborative design of services; easy government; child participation programme; LAAAB Space for democratic innovation; public policy co-creation processes [68].

Datos.gob.es. The Open data initiative of the Government of Spain is part of the Aporta Initiative to promote the opening of public information and development of advanced services based on data. The

Initiative aims to harmonize the synergies between ongoing open data projects, to drive and coordinate the actions of the different levels of the PA, the private sector and academia according to an integrated governance model, and to promote new products and services from the private sector and civil society. The portal currently includes 308 open data initiatives and 48,297 datasets in 75+ formats and 22 categories, such as: public sector; environment; economy; healthcare; science, transport, etc. [71]

Smart Service Platform Badajoz. Badajoz, the largest province in Spain, implemented a centralised smart platform providing tools and services based on IoT and Big Data to foster sustainable development, avoid rural exodus and increase the attractiveness of the region. The centralisation of all municipal services in one platform facilitates cost-efficiency and effectiveness. Badajoz aims to become the national reference for smart territory solutions and plans new pilots in parking, tourism, etc. [72].

2.2.2.3.1. The Autonomous Community of Aragon

Aragon Open Data Portal (AOD). AOD is the open data portal of the Aragon Regional Government (ARG), managed by the E-Administration and Information Society of the ARG. It relies on CKAN for the management of its open data system. AOD provides reusable and open data for citizens and businesses and an API REST (GA_OD_Core) with over 260 database views of different size, schema, and with heterogeneous data included in **Aragon's Open Data Catalogue**.

The **Aragon Open Data Pool** was developed in 2018 to connect and integrate disparate data sources from database views to the same schema and vocabulary: EI2A. **EI2A** describes basic data from the Government domain to support data interoperability and information standardization [69]. AOD has also carried out initiatives to bring the culture of open data closer to citizens, companies and all types of organisations, to encourage the development of services, and to facilitate data reuse in a simple way, e.g. via the **Open Social Aragon (OSA)** application. OSA captures and processes public information generated from the social network accounts of Aragon's institutions free social media APIs. It shows information on top hashtags, recent posts, heat maps, etc., and provides historic data [70].

2.2.2.4 Bulgaria

Strategy for the Development of E-government in Bulgaria 2020-25 (SDEBG). SDEBG stresses the importance of the use and implementation of innovative technologies such as open and Big Data, Cloud, AI, blockchain, IoT and role of PAs in opening large arrays of data that actually power technologies like AI, and introduces measures for their development [40]:

- Big Data is recognized as key for the development of e-services and innovations and SDEBG includes measures for the use of chatbots in eGovernment and the development of 3rd generation e-services using machine learning to predict user behaviour, etc. Another goal is the development of models for secure collection, storage, processing, use and reuse of non-personal data, dig data, high value data, etc., as well as the increased access to high quality data while ensuring that personal and sensitive data are protected
- Open data plays a major role in one of the priority projects – the registry reform, which requires the creation, accessing, storage and data management related to e-government, creation and

maintenance of data registers. According to SDEBG, Bulgaria is ahead of the EU average of 59% in the “Use of open data'' category and reaches 76%.

The Digital Transformation of Bulgaria 2020 - 2030 Strategy by MTITC sets the objective of deploying secure digital infrastructure, unlocking the potential of data and enhancing the efficiency of PAs and the quality of public services, the digital infrastructure, cybersecurity & data management [74].

"Bulgaria 2030" National Development Programme is a high-level strategic document determining the vision and general objectives of development policies in all sectors of the State, incl. digitalisation. Big and open data are a key part of the ***e-government*** sub-priority, incl. the development of high-speed connectivity, next generation infrastructure, data centres, Cloud infrastructure & digital inclusion [75].

Bulgarian Open Data Portal (BODP). BODP is a single, central, public web-based information system for the publication and management of reusable information in an open, machine-readable format together with the relevant metadata. BODP was launched in 2014 and currently includes 10,726 datasets from 533 organisations in 14 categories, e.g., environment, transport, governance and public sector, etc. In order to provide access to datasets that are not currently available, the portal includes an information request form [76]. Sofia Municipality also provides open datasets to be uploaded to BODP. [73]

2.2.2.4.1. Sofia

Projects set in DTSS and APDTSS. DTSS and the APDTSS recognize the importance and encourage the implementation of Big and open data solutions. One of the goals set in DTSS is the development of e-services based on open data and the promotion of the creation of a Big Data Centre of Excellence (CoE) via the creation of more open data repositories, data centres and open data policies, higher public engagement and data awareness, and collaboration with the local ecosystem to invent new solutions. Some of the key initiatives set in APDTSS to be implemented by 2023 include [43]:

- **Unified Platform for Urban Data:** creation of a Data Lake with structured and unstructured data in their original format for storage & use by PAs, businesses, citizens, research and educational institutions. The Platform will be expanded with modules for analysis, visualisation & modelling, creating a Data Warehouse. Time frame: 24 months (expected start of activities: H2 2021)
- **Development of a Unified Utilities Efficiency Model,** based on data on meteorological conditions and utilities consumption. Creation of an online platform for data collection and analysis to help utility companies increase resource efficiency and a database for meteorological conditions. Time frame: 24 months (expected start of activities: Q1 2022)
- **Register Reform & Data Policy:** review, update, consolidation and digitalisation of Municipal registers to ensure traceability, security and transparency. Development and implementation of a comprehensive data policy and a Roadmap for register digitalisation, upgrade and consolidation. Time frame: 18 months (expected start of activities: H2 2021)

- **Digital Twin of the City (collaboration with GATE Institute[4]):** creation of the first digital twin of Sofia to facilitate urban process- and decision-making optimisation via impact assessment, solution modelling and planning, R&D activities, etc. Pilot scenarios: urban planning, air quality & public transport in Lozenets district. Time frame: 24 months (expected start of activities: H2 2021)
- **Integrated Mobility Platform (IMP):** development of an open-data IMP providing real-time information on the modes of transportation and routes in the city to improve transport efficiency and provide open data to facilitate business creation and optimised Mobility-as-a-Service (MaaS) in Sofia. Time frame: 36 months (expected start of activities: H2 2021)

The Sandbox for Innovative Solutions is a programme of Innovative Sofia for testing innovations and identifying scalable pilot projects related to urban challenges (e.g. related to environment, transport, e-government, etc.). The projects are evaluated via an approved methodology developed with the support of the European Investment Bank and should be implemented in a short time – with a quick test and low financial risk. The Sandbox aims to generate data to support evidence-based city solutions and policies based on data; test innovative solutions and identify scalable projects, incl. related to Cloud, open and Big Data solutions; to increase the quality of public services and the quality of life in Sofia via innovation and smart urban solutions, etc. [77].

2.2.2.5 UK

UK National Data Strategy (NDS). NDS is an ambitious strategy aiming to drive the UK in building a world-leading data economy. Big Data is an important aspect in NDS and a key factor for the establishment of fairness, transparency and trust. The five priority missions of NDS include: unlocking the value of data across the economy, securing a pro-growth and trusted data regime, transforming government's use of data to drive efficiency and improve public services, ensuring the data security and resilience, and the international flow of data. Some of the concrete goals set in NDS include:

- the development of methods to use Big Data and modelling analyses to support a greater resilience of vulnerable countries to extreme weather events and disease outbreaks
- the empowerment of individuals and groups to control and understand how their data is used
- the promotion of fairness, transparency, data security and trust via application of new and emerging technologies, e.g. Big Data techniques and machine learning [78].

UK's Open Data Portal. Launched in 2010, it is a catalogue, making non-personal UK government data available as open data sets, currently covering 20K+ datasets in 14 categories from 1300+ national and local administrations. The catalogue is the delivery mechanism for the implementation of the INSPIRE geospatial directive and allows citizens to request data that has not been made public yet [79], [80].

[4] The Big Data for Smart Society Institute (GATE) is the only CoE in Big Data and AI in the Balkans and Eastern Europe, and was established in 2019 as an autonomous structure of Sofia University. For more information, please check the Official Website of GATE: https://gate-ai.eu/en/home/ .

The UK Government Licensing Framework (UKGLF) provides a policy and legal overview of the arrangements for licensing the use and re-use of public sector information, both in central government and the wider public sector. It sets best practices, standardises the licensing principles for government information, mandates the Open Government Licence (OGL) as the default licence for Crown bodies and recommends OGL for other public sector bodies [81].

Open Government Licence (OGL). OGL is a copyright licence developed by the National Archives as a tool to enable information providers in the public sector to licence the use and re-use of their Information under a common open licence and encourages public sector organisations to permit the use of their Information and databases under this licence. OGL is compatible with the Creative Commons Attribution License 4.0 and the Open Data Commons Attribution License, both of which license copyright and database rights [82] [5].

2.2.2.5.1. London, Camden

Open Data Camden (ODC) is a portal of the Camden Council, supporting the public, researchers and developers to access, analyse and share information about the borough. All data may be viewed, re-used and downloaded under an OGL Licence and the datasets through the main catalogue. ODC can be used for the creation of postings, visualisations and maps. All of the datasets are API enabled, allowing for the creation of new data-based solutions. The London dataset on Business & Economy, Children, Schools & Family, Community, Crime and Criminal justice and Education that is being used for London's use case in the PolicyCLOUD project can be accessed via the ODC [83].

The table below summarises the use cases and good practices described in Section 2:

Country	Examples of strategic documents	Examples of Projects/Good Practices
Italy	- National Innovation Plan 'Italia 2025' - Plan for IT in the PA (2019-21) - Cloud Strategy of the Public Administration - Italian Digital Government Code (DGC) - 4th National Open Government Action Plan 2019-21 under the OGP - Big Data Guidelines and Recommendations	- Cloud of the Public Administration model - Cloud marketplace - Cloud Enablement Program - #Data4All (SoldiPubblici, ItaliaSicura, OpenExpo) - Italian National Open Data Portal - Big Data in the Italian Institute of Statistics - Km4city
Spain	- Digital Transformation Plan for the General Administration and its Public Agencies	- Common e-Identification administrative system (eIDAS); - Cl@ve and @firma

[5] The UKGLF and OGL form part of the UK Government's drive to open up access to publicly held information and datasets, promoting transparency and enabling wider economic and social gain.

Country	Examples of strategic documents	Examples of Projects/Good Practices
	- National Plan to Stimulate the Demand for Cloud Solutions - Bill of Measures for the Implementation and Development of Cloud Computing in Aragon - 2020 Strategy for Social and Economic Recovery of Aragon -Digital Spain 2025 Strategy -4th Action Plan of Spain on the OGP	- Cloud-based recommendations system and Virtual desktops project (Aragon) - Aragon Open Data Portal, Open Data Pool, Open Data Catalogue - Open Social Aragon - Open data initiative of the Spanish Government - Smart Service Platform Badajoz
Bulgaria	- Strategy for the Development of E-government in Bulgaria 2019-25 - Digital Transformation Strategy for Sofia and the Action Plan for its Development - Strategy for Using Cloud Solutions in National Security 2014-20 -Digital Bulgaria 2025 National Program -Digital Transformation of Bulgaria 2020 - 2030 Strategy - Bulgaria 2030 Development Program	- State Hybrid Private Cloud (SHPC) - Cloud Electronic Signature - Unified Model (and Portal) for Access to E-Administrative Services of DAEU - MGovernment - With the amendments and additions to the Electronic Government Act adopted by the National Assembly, the State Agency for Electronic Government was closed and the Ministry of Electronic Government was established. By Decree № 12 of February 4, 2022 of the Council of Ministers, the Rules of Procedure of the Ministry of Electronic Government were adopted. SG no. 8 of 2022 - Innovative Sofia and School in the Cloud - Cloud E-services for the Administration - Bulgarian Open Data Portal - Sandbox for Innovative Solutions - Projects set in APDTSS
UK	- UK Government Cloud Strategy (G-Cloud) - Cloud Security Guidance - Government Cloud First Policy - UK National Data Strategy - UK Digital Marketplace Strategy	- UK's National Cyber Security Centre (NCSC) - GOV.UK Platform as a Service - Digital marketplace - UK's Open Data Portal - UK Government Licensing Framework - Open Government Licence - Open Data Camden
Germany	- Digital Strategy 2025	- Digital Twins - The Berlin Centre for Digital Transformation - Cloud Space Use

TABLE 5: CLOUD AND BIG DATA: GOOD PRACTICES FROM ITALY, SPAIN, BULGARIA, UK AND GERMANY

3 Data Lifecycle of Public Policies

A data lifecycle provides a high-level framework to plan, organise and manage all aspects of data during their life phases, from data planning to its destruction, and the relationship between phases [16]. Data lifecycle models are key factors for data management used by various public and private organisations, which helps them to provide exhaustive, readable and accessible data for diverse users. In this deliverable, we suggest a data lifecycle of public policies overview, adopting a Big Data Value Chain. The data lifecycle consists of four phases, shown in Figure 3 as follows: 1) Data acquisition and storage; 2) Data pre-processing, modelling, semantic representation and interoperability; 3) Analytics; 4) Visualisation [84], [85].

In addition, *data security, privacy and trust* are presented as a cross-cutting activity, arising during the whole data lifecycle [86]. The acquisition and storage phase have to follow security and privacy policies. Privacy-preserving data acquisition requires a method that does not infringe on privacy when collecting various data. This includes security mechanisms such as homomorphic encryption (method that can perform various operations without decrypting the ciphertext) and access control (ability to allow or deny someone to use something). The storage access control should be provided on physical as well as logical level. The pre-processing phase includes methods for ensuring data security and privacy such as encryption, anonymisation, etc. For example, De-Identification is a method of deleting the personally identifiable information in data or replacing it with attribute information. K-anonymity is one of the privacy-preserving techniques to prevent linkage attacks by linking public information and is used to prevent re-identification of de-identified privacy. The analytics phase includes privacy-preserving data mining and access control. The knowledge or patterns should be found implicit in data without infringing on the privacy of data owners. For example, association rule hiding is a method to prevent creating sensitive association rules in the analytics phase. The visualisation phase includes audit trail and privacy-preserving data publishing. The audit trail shows who uses the data, how and where it is appropriately used. The privacy-preserving data publishing delivers new unidentified or synthesized information that can be distributed to users without exposing the identity of the data subject.

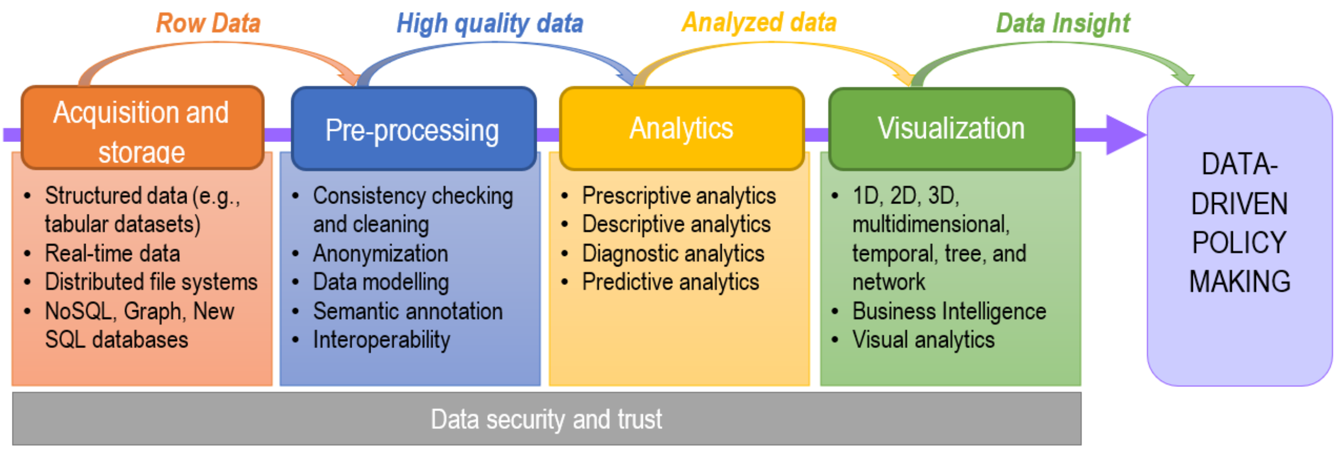


FIGURE 3: DATA LIFECYCLE OF PUBLIC POLICIES

The section below provides an overview of each of these phases, including definition and methods, description in the context of PolicyCLOUD, with a focus on the ethical, legal, regulatory, and societal considerations in the context of PolicyCLOUD as a key aspect of the public policies' data lifecycle.

3.1 Data Acquisition and Storage

3.1.1 Data Acquisition and Storage: Definition and Methods

Data acquisition includes data collection, structuring and storing in a scalable data storage. In order to achieve this, protocols that allow gathering information from distributed data sources of any type (unstructured, semi-structured, structured) and technologies that allow the persistent storage of the data retrieved, are required. Open frameworks and protocols for Big Data acquisition and the current approaches used for data acquisition in the different domains are presented in [87]. The commonly used open protocols for data acquisition are: AMQP (Advanced Message Queuing Protocol), supported by 23 companies became an OASIS standard in 2012; Java Message Service (JMS) is the de-facto standard for message passing in the Java world; the Memcached protocol is a memory caching system used by many sites (like Wikipedia, Flickr, Twitter, Youtube) to decrease database load, and hence speed-up the dynamic websites that work on the top of databases.

The following types of storage systems are being distinguished:

- Distributed File Systems e.g. the Hadoop File System (HDFS), offer the capability to store large amounts of unstructured data in a reliable way on commodity hardware. HDFS stores data as replicated blocks in files on the distributed file system providing fault-tolerance and high availability.
- NoSQL Databases use other data models than the relational model and according to the used data model they are distinguished to Key-value stores: Columnar Stores and Document databases.
- Graph databases such as GraphDB, Neo4J, Titan1 are persistent storage mechanisms that store all information as graph structures (nodes or edges), making them suitable for storing highly enabled associative or cognitive data processing of data, e.g. telecom networks, IoT, social networks and knowledge graphs. Knowledge graphs combine several important features: schema agility; logical inference; cognitive analytics based on network topology. A particular type of graph databases are RDF triple stores such as AllegroGraph, Virtuoso and GraphDB.
- NewSQL Databases are new forms of relational databases that aim at comparable scalability as NoSQL databases while maintaining the transactional properties (atomicity, consistency, isolation and durability) of traditional database systems.
- Big Data Querying Platforms are technologies that provide query facades in front of Big Data stores such as distributed file systems or NoSQL databases. Some representatives are Hive, Impala, Shark and Drill.

3.1.2 Data Acquisition and Storage in the Context of PolicyCLOUD

With the PolicyCLOUD framework, data acquisition can be done either for data streams or Ingest-now data. In both cases, the data flows through the PolicyCLOUD gateway and then through an Apache Kafka pipeline (with possible functions that may be applied along the pipeline) till data is ingested into the PolicyCLOUD data store. In PolicyCLOUD the Apache OpenWhisk platform is used to provide a Cloud environment supporting the FaaS paradigm. Apache OpenWhisk is an open-source, distributed serverless platform that executes functions in response to events at any scale. It manages the infrastructure, servers and scaling using Docker containers so one can focus on building efficient applications. The OpenWhisk platform supports a programming model in which developers write functional logic, called Actions, in any supported programming language, that can be dynamically scheduled and run-in response to associated events (via triggers) from external sources or from HTTP requests. OpenWhisk includes a REST API-based Command Line Interface along with other tooling to support packaging, catalogue services and many popular container deployment options. The use of OpenWhisk permits to register analytic ingestion functions which will be invoked by events triggered when data is ingested within the PolicyCLOUD framework. Examples of such functions are cleaning or sentiment analysis functions.

The data store used in PolicyCLOUD was till now the LeanXcale database (for details please refer to section 5.4 of the D4.3 deliverable [88]). As of 2021, PolicyCLOUD started adopting the “seamless” analytics framework from the H2020 BigDataStack project (detailed in section 4.4 of D4.3 [88]). It permits to store and analyse through SQL queries a dataset federated between a transactional database (in the occurrence, the LXS database) and an object store (only requirement is to support the s3 protocol). Both can be considered as a logical database that permits benefitting from both worlds: the operational and the analytical one. The seamless framework also keeps track of the current split point to show which part of a database is visible to a transaction in each table and seamlessly moves “historical” data slices from the LXS DB to the Object Store. The seamless framework is fully compliant with SQL, including JOIN queries between 2 datasets which are both distributed between the two underlying data stores. The JOIN implementation minimizes the amount of data transferred.

3.1.3 Data Acquisition and Storage: Ethical, Legal, Regulatory, and Societal Considerations in the PolicyCLOUD Context

As noted in PolicyCLOUD D3.6 applicable legal, regulatory, ethical and/or societal requirements may restrict an end-user’s leveraging of a data source which might otherwise be considered appropriate from a practical perspective (i.e., useful towards the goal which the end-user wishes to achieve). It is important to assess whether each of the data sources to be used can be exploited in compliance with such requirements (e.g., whether the data contained within the data source may be collected lawfully, considering the further processing within the platform and the end-goal for which it is collected by the end-user). As a result, specific data sources may need to be excluded or be restricted in terms of the amount or type of information extracted from them. In particular, the following key compliance controls defined in D3.6 [89] are relevant to data acquisition and storage:

- *“The platform and its end-users must weigh the interests of the data subjects appropriately and find effective means to provide information about the activities performed on personal data, considering also the need to preserve the quality of information in cases where providing this information may have an impact on the effectiveness of the use case.”*
- *“Data source quality must be controlled, by ensuring that only reliable sources are used, and to routinely test the analytics components of the platform to ensure that they do not skew knowledge obtained from data in a biased manner."*
- *"Should a selected data source be subject to contractual terms which prevent its use as intended by the PolicyCLOUD user, that data source should not be registered on PolicyCLOUD without proper authorisation from the data source owner.”*
- *“Should a selected data source be eligible for database copyright or sui generis right protection which prevents its use as intended by PolicyCLOUD, that data source should not be registered on PolicyCLOUD without proper authorisation from the rights holder.”*
- *“Should a selected data source, or a relevant part of that data source, which is to be extracted, be eligible for copyright which prevents its use as intended by the PolicyCLOUD user, that data source should not be registered on PolicyCLOUD without proper authorisation from the rights holder.”*
- *“PolicyCLOUD must assess which of the legal bases afforded by the GDPR may be applicable and implementable for an intended processing of personal data. This assessment must consider the full context of the processing activities which are intended, including the specific data sources to be used and the specific goals to be reached using the platform.”*
- *“PolicyCLOUD should only handle personal data in ways that may be reasonably expected and not use such data in a way that may produce unjustified adverse effects on data subjects.”*
- *“PolicyCLOUD and the end-users must be clear and honest with Data Subjects about the identity of the data controller which is collecting, processing and storing personal data, the methods used to process personal data, and the purposes of processing.”*
- *“PolicyCLOUD and the end-users, in relation to the processing activities which they may respectively perform, as controllers, shall lay down a specific and easily accessible document which duly informs Data Subjects of the processing activities carried out in the context of the Project: a privacy policy.”*
- *“PolicyCLOUD shall make available the Privacy Policy on its cloud-based platform, with appropriate steps taken to make it available to the Data Subjects whose personal data are used in the context of the Project. End-users should likewise ensure that the above information is available to Data Subjects on public websites under their control.”*
- *“PolicyCLOUD shall only collect personal data, which is adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. As such, for each purpose of processing connected to the Project, it shall identify the minimum amount of personal data needed to fulfil such purpose.”*
- *”Children inclusion into the investigation tool activities must be avoided, excluding collection of data related to minors, if possible, or storage after accidental collection.”*

- *"The end-user should identify steps to reasonably assure itself of the reliability of the data sources, in terms of the likelihood that any data collected from the data source may be false, inadequate, inaccurate or incomplete, also in terms of representativeness of the population of the borough, considering the purpose for which the data source is to be used. The end-user should identify, and document specific steps taken to address reliability and accuracy concerns detected."*

3.2 Data Pre-processing, Modelling, Semantic Representation and Interoperability

3.2.1 Data Pre-processing, Modelling, Semantic Representation and Interoperability: Definition and Methods

Data Pre-processing. Data pre-processing is related to data consistency checking and cleaning, including identification of the data records that are out of range, logically inconsistent or have extreme values, as well as processing of missing responses to minimise their adverse effects by assigning a suitable value or discarding them methodically. The data cleaning approach needs to be adjusted in terms of context and domain complexity, and according to the enhancement required to support the data analysis to be performed further.

Sample cleansing technique for e-commerce data is described in [90], in which crawlers are detected and regular deduping of customers and accounts is performed. In two other outlined cases, a probabilistic model for missing data in mobile environments and a system with application-defined global integrity constraints, to correct input data errors automatically are presented. Another example is a framework called BIOAJAX that standardises biological data for additional computation and eliminates errors and duplicates, so common data mining techniques can operate more effectively and searching quality is improved. Two other methods also discuss the robust and computationally efficient multivariate technique, called Minimum Covariance Determinant (MCD). The last computes the subset of h points from the data that minimises the covariance matrix determinant, which is very important for multivariate statistical methods, and Conditional Functional Dependencies (CFD), where the accuracy of data cleaning totally relies on the superiority of the dependencies used in data cleaning.

BIG project provides a very comprehensive analysis of the state-of-the-art, future requirements and emerging trends in the field of data curation, based on literature research, interviews with domain experts, surveys and case studies [87]. The main methods for data curation presented are:

- Master Data Management (MDM) that supports a single point of reference for the data of an organization and can be used to remove duplicates, standardize data syntax and as an authoritative source of master data.

- Collaboration Spaces such as Wikis (that scale to very large user bases) and Content Management Systems (that focus on smaller and more restricted groups to collaboratively edit and publish online content) allow users to collaboratively create and curate unstructured and structured data.
- Crowdsourcing is based on the so-called "wisdom of crowds" and advocates that potentially large groups of non-experts can solve complex problems usually considered to be solvable only by experts. The effectiveness of crowdsourcing has been demonstrated through websites like Wikipedia, Amazon Mechanical Turk, and Kaggle.
- Several other Data Curation Models such as Minimum information models, Curating Nanopublications: coping with the long tail of science and Investigation of theoretical principles and Domain Specific Models are also presented [87].

Data modelling, semantic representation and interoperability. In order to achieve a valuable result from data analysis, data should be embedded in a rich context during modelling. Unfortunately, the data is typically collected by institutions for internal purposes without attention on reusability, metadata and semantics. Furthermore, there are institutions that publish their data and metadata as open datasets, but without sharing data dictionaries, which can lead to misunderstanding and incorrect interpretation. Adding a semantic is a challenge, which requires connecting data back to its collection environment. It is a costly task, especially during the generation and acquisition phase of the data lifecycle. At the same time the ML algorithms are highly dependable on semantic information to learn from data precisely and unambiguously.

The efficient organization of data requires a knowledge organization system to be established. The glossaries, taxonomies, thesauri and ontologies are widely adopted NKOS that handle data ambiguity, provide hierarchical or associative relationships or data structuring in terms of classes, properties and constraints [91]. The semantic data annotations integrated with metadata allows computers to (semi-automatically) combine data from different sources [92] and enables interoperability. The analysis of semantics together with data allows new facts to be extracted and new enriched datasets to be generated. Two examples of good practices are being introduced below:

- **SMART DATA MODELS Initiative (SDMI).** SDMI is a joint collaboration between the FIWARE Foundation, TM Forum, and IUDX that aims to support the adoption of a reference architecture and compatible common data models to facilitate interoperable and replicable smart solutions in smart sectors, e.g. smart cities, environment, sensoring, health, energy, etc. A smart data model consists of 3 key elements: 1) the schema, or technical representation of the model defining the technical data types and structure; 2) the specification of a written document for human readers; 3) the examples of the payloads for NGSIv2 and NGSI-LD versions. All data models are public and royalty-free, granting the rights of free use, modification, sharing of the modifications, and free for all forever. The models are grouped into subjects, with every subject currently being a git repository, belonging to one or several domains (i.e. industrial sectors) [93].
- **CityGML:** CityGML is an open data model and XML-based format for virtual 3D city models storage and exchange, an application schema for the Geography Markup Language version 3.1.1 (GML3), and an international standard for spatial data exchange issued by the Open Geospatial Consortium

(OGC). Its aim is to reach a common definition of the basic entities, attributes, and relations of a 3D city model, especially the cost-effective sustainable maintenance of 3D city models, allowing the reuse of the same data in different application fields [94].

3.2.2 Data Pre-processing, Modelling, Semantic Representation and Interoperability in the Context of PolicyCLOUD

To formulate effective and properly structured evidence-based policies based upon diverse data, its pre-processing has to take place. Towards this direction, data cleaning techniques have to be applied in order to detect and correct (or remove) corrupt or inaccurate records from all the collected data that will be retrieved to formulate the required policies. In the context of PolicyCLOUD, such techniques are responsible for identifying all the incomplete, incorrect, inaccurate or irrelevant parts of all the collected data, and then replacing, modifying, or deleting the dirty or coarse data. Thus, possible missing, irregular, unnecessary or inconsistent data is found and totally cleaned. Especially dealing with missing data is one of the trickiest but common parts of the data cleaning process, since most of the models do not accept missing data.

To this context, the PolicyCLOUD cleaning techniques are able to: 1) identify errors associated with conformance to specific constraints, safeguarding that the data measures comply with predefined business rules or constraints; 2) correct/remove errors identified during the aforementioned validation process; 3) complete missing information, safeguarding that the data set provided is fully complete, and conforms to mandatory fields/required attributes (required fields which cannot be empty); and 4) safeguard that the data provided is accurate. To this end, a series of sub-mechanisms, orchestrated under the prism of the PolicyCLOUD Data Cleaning component, have been developed with the aim of undertaking the methodological processes aforementioned, taking into consideration the various types of analyses that need to be supported, assisting processes for filtering data according to use-case defined filters and rules, handling outliers' methods, handling missing values, as well as decomposing and smoothing data. The latter is being introduced in more detail in PolicyCLOUD Deliverable D4.3 [88].

However, having clean and qualified data is not the only prerequisite for formulating effective policies. Another key element in the policy-making domain addresses the scope of achieving and dealing with the very different formats and models of data that modern organizations need to process and analyse. Data interoperability addresses the ability of modern systems and mechanisms that create, exchange and consume data to have clear, shared expectations for the context, information and value of these divergent data. To this end, interoperability appears as the means for accomplishing the interlinking of information, systems, applications and ways of working with the wealth of data.

To address this challenge in the context of PolicyCLOUD project, the Enhanced Interoperability component introduces a novel approach that seeks to enhance the semantic and syntactic interoperability of the processed data, the SemAI framework. The latter is being introduced in more details in PolicyCLOUD Deliverable D4.3 and primarily focuses on the phases of the processing, the annotation, the mapping, as well as the transformation of the collected and cleaned data, which are the input to its system and which have major impact on the successful aggregation, analysis, and exploitation of data across the whole policy making lifecycle [88].

3.2.3 Data Pre-processing, Modelling, Semantic Representation and Interoperability: Ethical, Legal, Regulatory, and Societal Considerations

As noted in D3.6 [89], whenever personal data are collected from a registered data source for further processing on the platform, it is necessary, under the principle of accuracy, to ensure the accuracy and completeness of any personal data collected to the greatest extent possible. From an ethical/societal perspective, inaccurate or incomplete data may ultimately affect the output produced by the platform and used by end-users to create policies, which may lead to unforeseen and unjustified harmful impact on individuals and communities. In particular, the following key compliance controls, defined in D3.6 [89], are relevant to data pre-processing:

- *"Data source quality must be controlled, by ensuring that only reliable sources are used, and to routinely test the analytics components of the platform to ensure that they do not skew knowledge obtained from data in a biased manner."*
- *"PolicyCLOUD shall only collect personal data, which is adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. As such, for each purpose of processing connected to the Project, it shall identify the minimum amount of personal data needed to fulfil such purpose."*
- *"End-users and PolicyCLOUD must ensure that appropriate steps are taken to verify the accuracy of any personal data collected, to maintain those personal data up to date over time, and to allow data subjects to correct, complete or update their own personal data when needed."*
- *"PolicyCLOUD shall implement technical and organisational measures aimed at guaranteeing the accuracy and quality of personal data included in the cloud-based platform and shall provide means to Data Subjects for contributing to the maintenance of data that is always accurate and up-to-date."*
- *"PolicyCLOUD shall keep personal data in a form which permits identification of Data Subjects for no longer than is necessary for the purposes for which the personal data are processed. Even where personal data are collected in a fair and lawful manner, they cannot be stored for longer than actually needed, unless a reason for further processing exists, and provided that a legal basis for such further processing has been detected by PolicyCLOUD pursuant to the purpose limitation principle. Therefore, PolicyCLOUD shall proceed to the erasure of personal data from the cloud-based platform when it has no reasons for keeping them or, alternatively, it shall anonymize and aggregate such data."*

3.3 Analytics

3.3.1 Analytics: Definition and Methods

Four types of Big Data analytics exist as follows: Prescriptive Analytics, Predictive Analytics, Diagnostic Analytics and Descriptive Analytics [95]:

- Prescriptive Analytics aims to set up the structure, relationships and meaning of data;
- The Predictive Analytics forecasts the outcome using the collected data;
- The Diagnostic Analytics uses the past data to explain what and why something has happened;
- The Descriptive Analytics prescribes actions using the collected data.

Undoubtedly, Data Mining is the major technique used for predictive analytics. It identifies hidden patterns with the help of classification, regression, association rule and cluster analysis on huge datasets. It incorporates techniques from statistics and machine learning with database management. Machine Learning provides algorithms allowing automatically learning from data and enhancing through experience [96]. NLP is an example of a machine learning algorithm for human language analysis. One of its applications is using sentiment analysis on social media to obtain information about how customers react during product campaigns [95]. The supervised learning is based on the machine learning techniques that reason relationships using training data. In contrast, unsupervised learning identifies the hidden trends and patterns in data without using prior knowledge.

Ensemble Learning is a supervised learning technique providing better predictions than a simple constituent model. Classification is one of the most used supervised learning techniques. Taking as input data, it builds a classification model and maps a target class to the classifier. The Decision Tree is simple for implementation classifier, but due to its space limitation and overfitting problem it is not applicable for large data sets [96]. Another classification technique is the Support Vector Machine, but it has slow training and expensive computational time. The Neural Network algorithm has similar drawbacks due to its black-box nature and difficulty in interpretation. Its disadvantages are avoided by Deep Learning approaches providing enhanced performance. The main benefits of Deep Learning as analytics technique for Big Data are summarized in [97] as follows:

- Hierarchical layer appropriate for processing unstructured data;
- High-level abstraction providing complex representations of data make the machines independent of human knowledge;
- Process high volume of data based on a large dimensional raw data input;
- Universal model meaning that the physical models are used to characterized the universal phenomena;
- Does not overfit the training data while finding complex dependencies between different dimensions of the data.

Clustering is an unsupervised learning approach that classifies objects into groups of similar objects, whose similarity are not known in advance. One of the fastest clustering algorithms capable of handling large datasets are the Partitioning algorithms. K-means is a known partitioning clustering algorithm used in data mining and computer vision. Another popular clustering technique is provided by the graph-partitioning-based algorithm that is appropriate for partitioning a graph G into subcomponents with specific properties. The second group of clustering methods is presented by the Hierarchical algorithms that follow two main approaches, namely Agglomerative (bottom-up) and Divisive (top-down). Density-based algorithms form a third group of clustering algorithms that search for clusters of nonlinear and arbitrary shapes. They are widely adopted in medical datasets such as biomedical images. In contrast to the classification techniques that are based on a discrete target, the regression techniques use a continuous target. Their main purpose is to identify how the value of an independent variable is changed according to the value of one or more independent variables. Such techniques are suitable for predicting sales volumes taking into account different economic markets and variables.

Evolutionary Computation is recognized as a new technique for finding optimal solutions and solving real world problems [98]. Since it enhances the performance of information retrieval, the Evolutionary algorithms are applicable in the field of Big Data. There are four main methods of Evolutionary Computation as follows: Genetic Algorithms, Evolutionary Strategies, Evolutionary Programming and Genetic Programming. Such methods are suitable for solving nonlinear problems, such as improving job scheduling in manufacturing and optimizing the performance of an investment portfolio [95].

3.3.2 Analytics in the PolicyCLOUD Context

The PolicyCLOUD framework permits to both register datasets and analytics functions through the Data Acquisition and Analytics Layer (please see section 2 of the D4.3 deliverable for details [88]). The registered analytics functions may be applied through the PDT either on streaming data (that is on a stream of data that pertains to a given dataset) or on a dataset after it was ingested within the PolicyCLOUD data storage. This later case will be further described in this subsection.

It should be remarked that when both the dataset and the analytics function where registered the registrant party has to fill-in documentation concerning legal/privacy aspect pertaining to the dataset (e.g., bias documentation, GDPR documentation, licensing documentation, etc.) or to the analytics function (e.g., trade-offs documentation, bias documentation, etc.) so that the policy analyst/make party who is applying the function on (part of) the dataset has the occasion of checking the legality of this action. A dataset/function that is lacking relevant documentation should be considered as uncertain and therefore probably not be used. For the case of PolicyCLOUD, the expected analytics functions mostly will fill in the following categories:

Exploratory data analysis (EDA): For most of the use cases the exploration of one variable (or a combination of them) by the policy maker become the starting point for policies assessments and set the basic for further and more complex analysis as cited in [99]: "The more one knows about data, the more efficiently it can be used to develop, test and refine a theory". Some of the techniques used for EDA are statistical summaries (numerical summaries of the data) or visual representations (visual displays of data). Considering that the emphasis of EDA is on using visual displays to reveal vital information about the data being examined [99], most of the EDA analytic functions in PolicyCLOUD will rely on a graphical charts-based technique for examining the variables & relationships between them.

Sentiment and opinion mining: Internet and especially social media have become one of the most significant data sources for knowing the opinion of citizens regarding multitude aspects or topics. This growth of social media opinions has raised the need of researching in multiple fields (NLP, machine learning, data and text mining, etc.) in order to come across the opinion and sentiment from these publicly available data sources [100]. For the case of PolicyCLOUD extracting the sentiment of a piece of text become essential for most of the use cases, primary example being use case 2 “Intelligent policies for agri-food industry”, scenario B, interested in knowing negative and positive opinions of the different wine products analysed on social networks.

Predictive analysis, one of the branches in advanced analytics, is the capability of forecasting activities, behaviour, or trends and in general unknown results. This technique in its essence tries “to discover the relationship between independent variables and relationship between dependent and independent variables” [101] and encompasses multiple disciplines as statistical analysis, machine learning and operations research [102]. As part of the predictive analysis, “Trend Analysis” is the capability of comparing historical data over time for identifying any solid or trend results. This analysis might as well result significantly useful for the PolicyCLOUD use cases. In this regard, deep research in both pure trend analysis definitions, such as that proposed by Harvey in [103] or novels definitions more focused on social media [104] will be done with the aim of helping the use cases in the identification of trends within the data of their interest.

3.3.3 Analytics: Ethical, Legal, Regulatory, and Societal Considerations

As noted in D3.6 it should be ensured that enough testing is performed to ensure a reasonable degree of statistical accuracy for the analytics activities. Any operations performed on data should be methodically logged to ensure traceability, and the general rationale and logic behind the analytics performed should be explained to end-users, so this can be considered during their decision-making process. End-users should be informed of the possibility of false positives or false negatives and errors in the presentation of the results, so that they are incentivised to critically examine results produced by the analytics functions in their decision-making process. In particular, the following key compliance controls, defined in D3.6 [89], are relevant to analytics:

- “Data source quality must be controlled, by ensuring that only reliable sources are used, and to routinely test the analytics components of the platform to ensure that they do not skew knowledge obtained from data in a biased manner."
- “Regarding the Enhanced Interoperability component (EIC), PolicyCLOUD should ensure that enough testing is performed to ensure a reasonable degree of statistical accuracy for annotations and connections established. This may require an extensive training exercise involving various kinds of data sources, to refine rules used by this component for these activities.”

- “The Enhanced Interoperability component shall methodically log the activities performed to ensure traceability and present correlations established and their rationale to the end-user for confirmation, to provide some level of human validation of those correlations. End-users should also be advised of the possibility of false positive correlations, so that they are incentivised to verify the validity of correlations made.”
- “PolicyCLOUD should ensure that enough testing is performed to ensure a reasonable degree of statistical accuracy for the analytics functions. Any operations performed on data should be methodically logged to ensure traceability, and the general rationale and logic behind the analytics performed should be explained to end-users, so this can be considered during their decision-making process. End-users should be advised of the possibility of false positives or negatives and errors in result presentation, so that they are incentivised to critically examine results produced by the analytics functions in their decision-making process.”

3.4 Visualisation

3.4.1 Visualisation: Definition and Methods

Data insight provides both accessibility and proper visualization, including advanced visualization techniques that consider a variety of Big Data (i.e. graphs, geospatial, sensor, mobile, etc.) available from diverse domains. The most used and spread visualization tools and techniques for large data sets focusing on their main functional and non-functional characteristics are surveyed in [105]. The surveyed tools are grouped in 4 subgroups: 1) data visualization (whose main goal is to communicate information clearly and efficiently to users, involving the creation and study of the visual representation of data); 2) information visualization (whose main task is the study of (interactive) visual representations of abstract data (numerical and non-numerical data, such as text and geographic information) to reinforce human cognition); 3) scientific visualization (for spatial representation) and 4) business intelligent and visualization tools.

The commercial Visual Analytics frameworks are explored in [106], complementary to an existing survey on open-source visual analytics tools. The visualization techniques are divided into graphical representations of data and interaction techniques. On a high level, they are classified by the type of visualized data: 1) numerical data (bar chart, line chart, pie chart and scatter plots, parallel coordinates, heatmaps, and scatter plot matrix); 2) text/web (word cloud and theme river); 3) geo-related data (projection on map); and 4) network data (graph) (Treemap, other graphs).

Techniques such as 1D, 2D, 3D, multidimensional, temporal, tree, and network are important to get deep insights about the large volume of data. Among different types of chart visualisation, the bar chart type has better usage compared to the pie and line type chart. Visualisation through maps type is used to analyse multidimensional types of data. Heat types of maps are widely used by the companies to analyse the data in an effective way, followed by Geo, Tree and Spatial. Also, there are other miscellaneous tools used by Big Data to analyse the data like tables, spreadsheets, counts, data sheets, histograms, scatterplots, and statistics.

3.4.2 Visualisation in the PolicyCLOUD Context

Focusing on the PolicyCLOUD platform, visualisation, considered as the last step of the data lifecycle, aims to help policy makers understand how the policies they are monitoring evolve, considering the selected KPIs for each policy and the data that populates the PolicyCLOUD data store in that moment. Each chart, designed following the use case needs, will display the output of the PolicyCLOUD analytics to provide the policy maker with an intuitive way to read and understand these analytics outputs and facilitate the extraction of conclusions from them, and then use it to improve their further decisions. As extensively detailed in deliverables *D5.2 Cross-Sector Policy Lifecycle Management: Design and Open Specification 1* [107] and *D5.4 Cross-Sector Policy Lifecycle Management: Design and Open Specification 2* [108], different types of visualisations that can be used to display the results of the analytics. Each type of chart has its own characteristics that make it more suitable for each type of data. This suitability between data type and charts fixes which ones fit better for each scenario. Some examples of how charts, through the results of the analytics component, can help and inspire policy makers are:

- MAGGIOLI: Among other charts, this use case uses a heat map to reflect the results of radicalization analytics. This map provides a quick vision of radicalization hotspots, allowing actions to be taken by policy makers, in less time.
- SARGA: In this use case, among other charts, a Gauge can be found, through which policy makers can get a quick idea of the average sentiment obtained for a certain topic. In the Gauge, the average opinion/sentiment is shown with a value between -1 (very negative) and 1 (very positive), allowing conclusions to be drawn quickly by policy makers.
- SOFIA: As an example, for the Sofia use case, a horizontal bar chart displays the results of the analytics for the Road Infrastructure scenario. From this chart, it's possible to get a quick idea about the road-related incidents in the 24 districts of Sofia Municipality, allowing policy makers to take appropriate actions.
- LONDON: For the London use case, one example could be a line graph for its scenario of predictive analysis. In this case, the line graph helps policy makers make some future outcome predictions, based on the results of the analytics shown in the chart.

These are just some examples of the whole set of charts, agreed with both, the pilots and the analytics component partners, chosen to be the most appropriate ones, in each case, to achieve the proposed goals according to the output expected by the pilot partners (the output of the visualisations) and the input of the visualisation (what the analytics component use to feed the visualisation component).

3.4.3 Visualisation: Ethical, Legal, Regulatory, and Societal Considerations in the PolicyCLOUD Context

As noted in D3.6 [89], the data visualisation components should ideally not actually rely on or disclose information about identifiable individuals, but instead rely on appropriately aggregated data, which cannot be traced back to any given specific and identified individual, to avoid raising personal data protection concerns regarding data manipulated by PDT users. Furthermore, another key concern around the means used to visualise data, as well as insights derived from such data, is ensuring that visualisations are understandable to the platform's end-users. This means not only ensuring that the actual output presented by the platform is understandable (in that it allows the average end-user to reasonably derive relevant information from it), but also that the process followed to reach that output is carefully explained (so that the average end-user is able to understand the operations performed by platform on data to reach the output).

It is important that end-users are able to understand the operations performed by the platform on the data; otherwise, they may not be able to understand the connections established by the platform. Where this understanding is missing, the end-users may not feel confident in relying on the output of the platform or may be deprived of any ability to critically examine that output and ensure that it amounts to a correct interpretation of the context in which the data are being considered. End-users should also be advised of the possibility of false positive correlations, so they are incentivised to verify the validity of correlations made. Input from end-users can be used to further refine correlation-defining rules. This ability to explain how the platform produces certain results, and the results themselves, is also important to ensure that end-users fully understand and critically examine the results shown to them. This should, in turn, allow those end-users to use those results to make reasoned policy-making decisions, which they can, in turn, explain to the affected individuals. In particular, the following key compliance controls, defined in D3.6 [89], are relevant to visualisation:

- *"Mechanisms facilitating the auditability of AI systems (e.g., traceability of the development process, the sourcing of training data, and the logging of the AI system processes, outcomes, and positive and negative impacts) should be put in place."*
- *"Any trade-off between requirements, principles or individual rights considered in AI system development should be properly documented."*
- *"The Enhanced Interoperability component shall methodically log the activities performed to ensure traceability and present correlations established and their rationale to the end-user for confirmation, to provide some level of human validation... End-users should also be advised of the possibility of false positive correlations, so thatthey are incentivised to verify the validity of correlations made."*
- *"*PolicyCLOUD *should ensure that enough testing is performed to ensure a reasonable degree of statistical accuracy for the analytics functions. Any operations performed on data should be methodically logged to ensure traceability, and the general rationale and logic behind the analytics performed should be explained to end-users, so this can be considered during their decision-making process. End-users should be advised of the possibility of false positives or false negatives and errors*

in result presentation, so that they are incentivised to critically examine results produced by the analytics functions in their decision-making process."

- *"End-users shall be enabled to understand the output presented to them by the PDT."*

3.5 Data Protection and Privacy Enforcement

3.5.1 Data Protection and Privacy Enforcement in the PolicyCLOUD Context

As mentioned in the introduction to Section 3, data security, privacy and trust is presented as a cross-cutting activity, arising during the whole data lifecycle [86]. PolicyCLOUD facilitates data protection and privacy enforcement, through authorization and authentication mechanisms that provide logical access control to generated data. *Authentication* deals with the problem of proving the identity of a natural person or a system entity that aims to interact with a resource while *authorization* deals with the problem of providing logical access control to these resources. The logical access control is always bound to specific actions that the natural person/systemic entity (hereinafter subject) wishes to perform i.e., discovering, reading, creating, editing, deleting, and executing.

Regarding authentication, Keycloak is used as an authentication provider and authentication proxy for micro-services and legacy systems. As such, it abstracts the functionality of identity extraction and identity verification for different systems and for different protocols. In parallel, it is able to map users and roles from existing legacy systems in what it calls authentication realms.

Regarding authorization, there are several models and mechanisms, with each having its own advantages and limitations. The most dominant Access Control Mechanism are listed below:

- **Discretionary Access Control (DAC)** where the owner of the object specifies which subjects can access the object. Most operating systems such as Windows, Linux, Macintosh and most flavours 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 scheme, according to the XACML standard, **makes it ideal for protecting the data in the lifecycle of PolicyCLOUD**, as it allows dynamicity in policy creation and also separation of concerns among policy definition and policy enforcement.

3.5.2 Data Protection and Privacy Enforcement: Ethical, Legal, Regulatory, and Societal Considerations

As noted in D3.6 [89], it is essential for the PolicyCLOUD platform to incorporate adequate technical and organisational security measures, developed as a result of a dedicated security risk assessment targeting potential threats generated, in particular, from reliance on Big Data analysis. This is not only relevant from the legal/regulatory perspective – particularly where personal data may be processed via the platform – but also from the ethical/societal perspective, as technical robustness and safety has been identified as one of the key requirements for trustworthy AI by the High-Level Expert Group on Artificial Intelligence [109]. Compliance with relevant cybersecurity certification schemes (e.g., the European Union Agency for Cybersecurity's draft European Cybersecurity Certification Scheme for Cloud Services) [110] may prove a practical manner in which to ensure a high level of security for the platform. In particular, the following key compliance controls, defined in D3.6 [89], are relevant to data protection and privacy enforcement:

- *"The platform should incorporate adequate technical and organisational security measures, developed as a result of a dedicated security risk assessment targeting potential threats generated in particular from reliance on Big Data analysis PolicyCLOUD should implement technical measures, including hashing and cryptography, to limit the possibility of repurposing personal data."*
- "*Platform users shall be limited both from a technical and from a contractual point of view in how they can process personal data which are collected and managed through the* PolicyCLOUD *platform.*"
- "*It should be demonstrable that the security measures implemented on the platform were chosen as a result of a documented risk assessment, with justifications as to why those measures were deemed adequate to address the specific risks identified.*"
- "*There should be clearly defined rules and specific channels on the reporting of security incidents or abnormal events related to the platform., and all persons working with the platform should be made aware of the types of occurrences which may qualify as a reportable security incident.*"
- "*The processes implemented to address personal data breaches by* PolicyCLOUD *should ensure that all relevant information on a personal data breach and the manner in which it was handled is documented in a register of personal data breaches, as set out in Art. 33, par. 5 GDPR […].Post-breach analyses should also be carried out, to validate the effectiveness of the breach management process, identify areas of improvement, and identify, based on a root cause analysis*

of the incident, adequate technical and organisational measures to reduce or eliminate the likelihood of recurrence."

- "*A security risk assessment, as part of an overall DPIA, should be carried out, to identify possible threats and risks to the fundamental rights, freedoms and interests of Data Subjects and the specific security measures implemented or which should be implemented to address them."*
- "PolicyCLOUD *shall perform a comprehensive risk assessment, focused on the likelihood and impact of threats to confidentiality, integrity and/or availability of assets stored on the platform, and to the resilience of the infrastructure of the platform and systems itself, as well as relevant compliance obstacles raised by use of the Cloud infrastructure, and assess whether the technical and organisational measures put in place by the IaaS provider in relation to the Cloud infrastructure sufficiently mitigate any relevant risks identified."*

3.6 Data Space Ecosystems in the Public Sector

The concepts of data spaces and data space ecosystems are another key element in the context of the lifecycle of public policies. Data spaces are a data management abstraction that aims to solve some of the issues with data integration systems. A decentralized data ecosystem created by the parties with a common goal is called a data space. It can also be described as a decentralized data ecosystem built on widely accepted building pieces that enables reliable and efficient data sharing among members. An ecosystem with a governance structure is formed by data consumers, data producers, and identification and authorization service providers to build a fundamental data space. In order to accommodate a range of use cases, service providers such as marketplaces, brokers, billing and clearing services, etc., can be included in the data space. There are currently 2 different public initiatives that aim to lead the creation of these data spaces The International Data Spaces Association (IDSA) [113] and GAIA-X [114] .

The International Data Spaces Association (IDSA) is working to develop International Data Spaces (IDS), a secure, sovereign system of data sharing that allows all users to realise the full value of their data. IDS ensures that data providers retain self-determined control over how their data is used (see Figure 4) or "data sovereignty," while enabling new "smart services" and creative business processes to operate across organizations and industries.

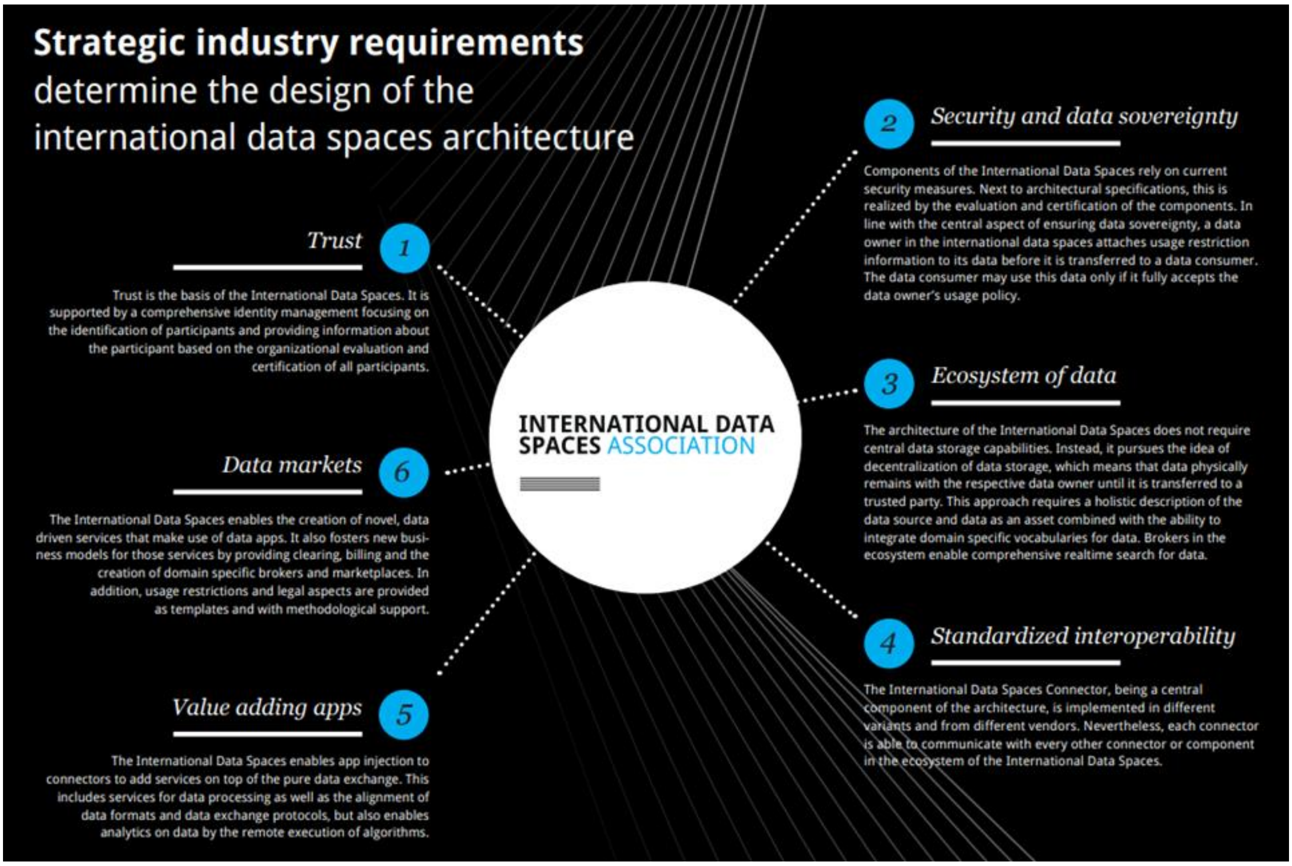


FIGURE 4: STRATEGIC INDUSTRY REQUIREMENTS FROM IDS. SOURCE: DATA SPACES SURVEY MAY 2022

The Gaia-X European Association for Data and Cloud AISBL represents the core of the organisational structure. It is an international non-profit association under Belgian law (French: association internationale sans but lucratif, shortened to AISBL). It was founded to develop the technical framework and operate the Gaia-X Federation services. Officially, the Association was founded by 22 companies and organisations in January 2021. Until today, over 340 members have joined and more are welcome. Its members are committed to upholding the values of data protection, transparency, openness, security, and respect for data rights

Gaia-X Vertical Ecosystems support the creation and development of data spaces and projects. The aim of vertical ecosystems is to share data space knowledge and collect cross-country use-cases while giving Gaia-X Association members the opportunity to network, collaborate and identify open standards related to specific domains. Gaia-X Data Space Business Committee supports the link with and across the vertical ecosystems and ensures the development and application of Gaia-X deliverables in specific domains. An indicative list of the current vertical ecosystems includes the following: Aerospace, Agriculture, Circular Economy, Education, Energy, Finance, Geoinformation, Health, Manufacturing / Industry 4.0, Media, Mobility, Public Sector, Smart Cities, Smart Living, Space and Logistics

Currently work is being done to integrate both IDS and GAIA-X architectures, gaining the benefits of both infrastructures. The GAIA-X and IDS integrated architecture supports and allows data spaces and creates cutting-edge smart services across industry verticals. IDS focuses on data and data sovereignty, whereas GAIA-X focuses on sovereign cloud services and cloud infrastructure. Self-sovereign data storage, reliable data use, and interoperable data interchange are the three key goals of the relationship between GAIAX and IDS. In this approach, GAIA-X promotes smart data applications and innovations across industry sectors while being created in compliance with the European Data Strategy. In order to provide cloud and data sovereignty for whole data value chains in federated ecosystems, GAIA-X and IDS complement one another.

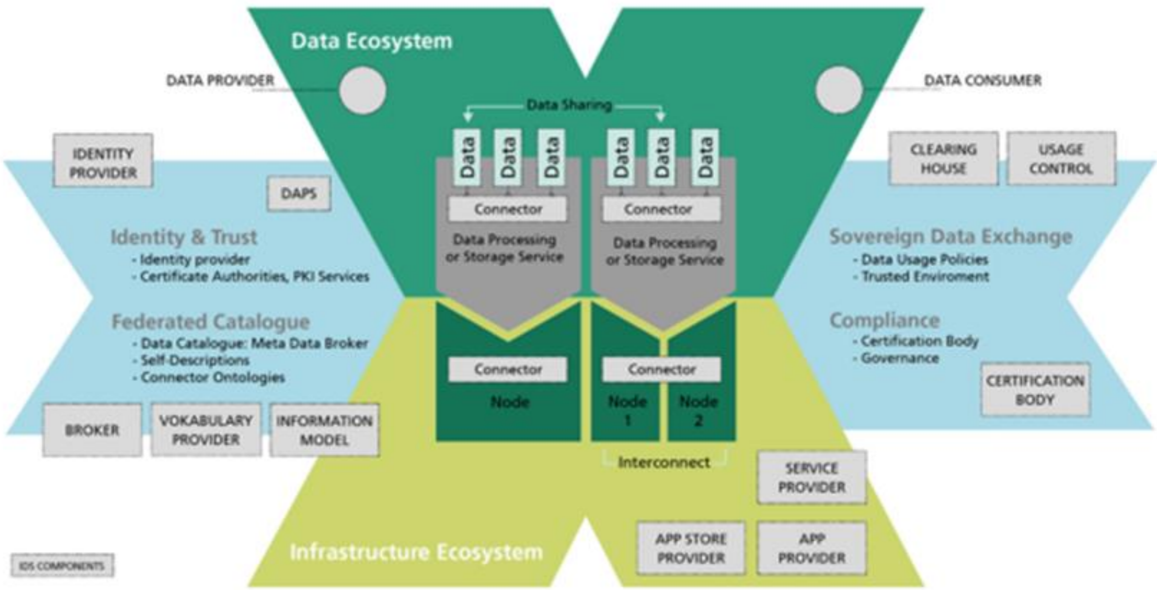


FIGURE 5: MAPPING OF IDS COMPONENTS INTO THE GAIA-X ARCHITECTURE (SOURCE: GAIA-X INITIATIVE)

Currently there are two large European public projects that intend to work on the generation of these data spaces that are benefiting from both initiatives such as Gaia-X and IDS, these two projects are i4Trust and EUHUBS4DATA. Where different PolicyCLOUD partners are involved. i4Trust is building a sustainable ecosystem where companies will be able to create innovative services by means of breaking "data silos" through sharing, re-using and trading of data assets. On the other hand, EUHubs4Data project is building a European federation of Data Innovation Hubs based on existing key players in this area and connecting with data incubators and platforms, SME networks, AI communities, skills and training organisations and open data repositories.

These initiatives allow the generation of digital ecosystems that allow to bring together both data and new services for different sectors, being one of these sectors the public sector that can be both data provider and data consumer for the construction of cloud services for decision making. PolicyCLOUD is no stranger to these initiatives as they turn out to be one more element in the data cycle.

4 Methodology for Policy Making in Cloud Environment

For the purpose of this deliverable, the term *methodology* is defined as "a set of methods, tools and rules of elaborating, monitoring and evaluating a given policy." *Method* can be defined as: "a procedure, technique, or way of doing something". While the *methodology*[6] is more general, the *method*[7] is more technical in nature than the whole presented by the methodology.

4.1 Methodology Elaboration/Development

The PolicyCLOUD methodology adapts the **Kaizen methodology** [111] and the concept of **digital twins**. The Kaizen methodology prescribes the continuous improvement in an organization, based on the constant small positive changes that can result in major and more significant growth. The concept of digital twins is not new, since it has been applied in computer-aided design (CAD) for over 30 years. However, currently there is a new view on it, and the concept was listed by Gartner as one of the top 10 strategic technology trends that guides digital transformation. "Urban digital twins can change the way cities are planned, operated, monitored and managed" [112].

The methodology has two main inputs: knowledge and data. Although knowledge can be extracted from data, the proposed methodology assumes that knowledge comes from the expertise of the PAs. As part of the suggested methodology, a set of adaptable approaches for the examination of public authorities' policy-making structures and their systems maturity to absorb and analyse data in common Cloud infrastructures will be further developed in the second iteration of D 5.1 in M32 and added to the first step of the methodology 'problem definition'. The analysis of best practices in Section 2.2 of this deliverable can be seen as a first step towards the development of those approaches.

Problem definition: problems are identified, the maturity assessment of the public administration is conducted, and the need for action is formulated in a given domain. Maturity will be assessed by such categories as political will, presence of public institutions in charge of digitalisation and innovation, including Cloud and Big Data development, availability of strategic documents, and concrete infrastructures, projects and initiatives related to Cloud and Big Data implementation in the public sector. As a result of the problem definition, policy context will be established.

Policy discussion: The right approach to meet the problem, defined on a previous stage, is identified, and KPIs for measuring the effectiveness of the approach are defined. All factors affecting city performance, which are related to it, should be analysed in order to have the best chance for policy success. As a result of the policy discussion the actual policy will be formulated together with an objective way of its evaluation through KPIs. All stakeholders involved in the policy and other actors will also be identified.

[6] Methodology comes from Greek – "method" and "logos".

[7] Method comes from Greek – "metha" (through) + "hodos" (way)

Policy harness: The available data for problem analysis and KPIs calculation is collected. For each KPI, a calculation method is defined. The data could be gathered from different sources, such as sensors, cameras, open data, social media, and other systems that serve the needs of public authorities, enterprises, utility companies providing water, electricity, natural gas and other amenities, etc.

The following 6 data types have been identified, based on the fact that they often lead to the use of different approaches and techniques for data analytics and data storage: (1) Structured data; (2) Time series data; (3) Geospatial data; (4) Media, Image, Video and Audio data; (5) Text data, including Natural Language Processing data and Genomics representations; and (6) Graph data, Network/Web data.

At this stage, data ingestion into a data storage and data pre-processing are performed. Data ingestion could be performed in two ways: (1) by importing datasets in different formats such as .csv, .xlsx, .xml, .json, and (2) by consuming external APIs, provided by PAs. The second one is preferable since it allows continuous integration of data and more thorough analytics. The APIs enable easier data integration from different public authorities and the development of complex analytical models that generate additional revenue. For example, cities or cities' departments working on similar problems could continuously combine their data to obtain better decision support, since more data ensures more dimensions to explore, better training of analytical models, wider purview and more detailed results.

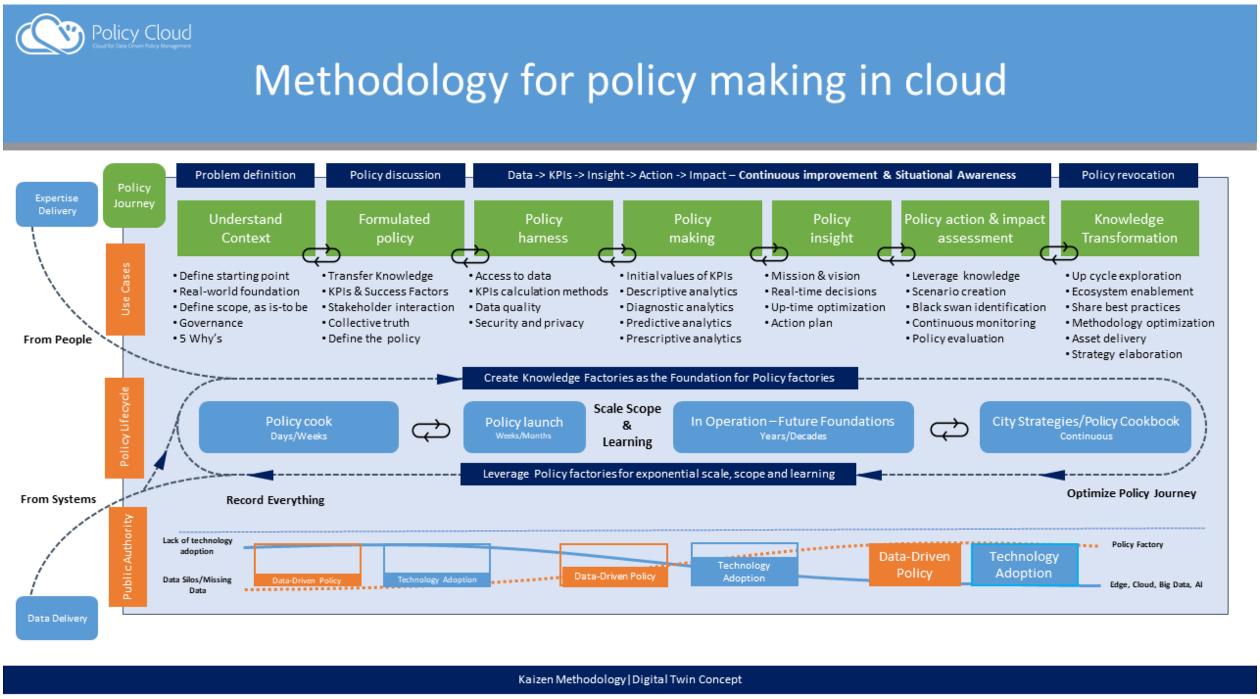


FIGURE 6: POLICYCLOUD METHODOLOGY FOR POLICY MAKING IN CLOUD ENVIRONMENTS[8]

Data pre-processing includes data cleaning, consistency checking, linking, anonymization and other techniques aiming to ensure data quality. This avoids the so-called GIGO (garbage in, garbage out) situations, since the quality of output is determined by the quality of the input. In other words, if the data

[8] The full-sized methodology can be found in Appendix A

used in analyses on the next stage is not reliable or coherent, the results will not prove useful. The pre-processing activities could be performed on the premises or in the Cloud. Another important aspect that should be considered is related to data security and privacy.

Regarding the approach of calculation, the KPIs can be divided into single, aggregated, composite and indexes:

- The single KPIs deal with a single dimension that is being measured in a given unit. The concentration of CO2 emissions and the number of private cars available in a city are examples of single KPIs.
- The aggregated KPIs are based on two or more dimensions that are being measured in the same unit. An example of aggregated KPI is the intensity of CO2 emissions by type of source (production, transportation and residential users) or by sector.
- The composite KPIs compile several features of a given dimension of the city/region/country. A single value in a common unit is obtained as a result. For example, the composite KPI for measuring the nuisance is calculated as a percentage of the population affected by noise, odour and visual pollution.
- The indexes are KPIs that aggregate several dimensions of the smart city in a single unitless value. The Human Development Index (HDI), for example, is a summary measure of average achievement in three dimensions of human development related to health, education and standard of living.

Policy making: The main goal is to produce insights that drive the decision-making process. Analytical approaches and algorithms based on machine learning, sentiment analysis and other Artificial Intelligence methods have to be applied. Additionally, the initial values of the KPIs are computed based on the calculation formulas defined on the previous stage. The analytics could be performed in different forms:

- Descriptive Analytics shows what happened in the past and help understand the current situation, e.g. in the form of data visualizations like graphs, charts, reports, and dashboards.
- Diagnostic Analytics answers why something happened in the past. The diagnostic analysis is often performed through root cause analysis. This includes application of techniques such as data discovery, drill down and drill through, data mining.
- Predictive Analytics predicts what is most likely to happen in the future. Machine learning models are usually created to perform predictive analytics. As a result, key trends and patterns could be identified.
- Prescriptive Analytics recommends actions that can be taken to change the future. It suggests different directions of action and gives evidence what the potential implications for each of them would be.

Effective data management and data analytics might imply the effective usage of Cloud and HPC infrastructures, especially in the case of Big Data.

Policy insight: The insights from analytics are visualized in dashboards, including a variety of charts and summary reports. Their aim here is to show areas where improvements and further investigations are potentially needed. Decisions on how to implement the policy in the most efficient way could be taken, and planning for future implementation needs to be elaborated, including region of intervention, time constraints, etc.

Policy action & impact assessment: The successful implementation of the policy requires scenario creation and planning black swan identification. During the implementation of the policy, its impact needs to be monitored. The process itself could also be influenced by data in two directions.

- First, different levels of policy intensity could be implemented (pinpoint problem zones). For example, a decrease in traffic can be focused more precisely on problem areas, thereby reducing the occurrence of congestion at the point of its origin.
- Second, the execution of the policy produces new data, which can then be used to evaluate its effectiveness through KPIs. Thus, the future implementation of the policy could be enhanced and policy refinement could be made.

Knowledge Transformation. No matter how successful a policy is, the results of its implementation need to be recorded in order to have a solid knowledge base for future planning. Successful policies are shared as best practices and might be used for development of strategies.

4.2 Prerequisites for Usage and Applicability

In order to unlock the full potential and benefits of the methodology, public authorities must meet the following **prerequisites**:

- a clear overall vision on what is the final goal to achieve;
- decision by the policy maker(s) to employ the methodology and the underlying platform;
- readiness to work together on interdisciplinary, cross-sectoral issues;
- efficient communication and information flow;
- availability and accessibility of data, considering the growing data flows;
- ability to translate data into valuable assets;
- ability to specify KPIs, which can be easily interpreted and used in daily operations;
- willingness and ability to co-create.

The methodology could be **used** for a variety of purposes as follows:

- measuring performance of policies (before and after situations) and public authorities' internal self-assessment;
- comparing the expected impact with a reference situation;
- supporting the development of new approaches to public services by making data accessible and reusable;
- as a support tool during a strategy setting process;
- monitoring the progress of the public authority as a whole towards public strategy;
- identifying tasks that could be performed in a better, more effective and efficient way;
- benchmarking of policies to find differences and similarities;
- sharing best practices between the public authorities in order to learn from each other;

- assessment of the level to which projects meet the goals set at the beginning;
- facilitating and inspiring the learning process;
- external (independent) assessment.

The following **decision-making processes** could be supported by the methodology:

- strategic decisions;
- performance-based budgeting;
- translating strategic goals into actions;
- checking the performance at small scale and planning future expansion.

4.3 Methodology Validation

The PolicyCLOUD platform supports the application of the methodology in order to create a particular policy or several policies. This section describes the validation of the methodology using the use case of the City of Sofia as an example. The feasibility of the methodology is proved by showing how it can be applied following the policy generation steps provided by the PolicyCLOUD platform.

4.3.1 Mapping Between the Methodology and PolicyCLOUD Platform

Table below 6 provides a mapping between the stages of the methodology and the steps for policy generation that are implemented in the PolicyCLOUD platform. It provides a base for further validation of the methodology presented in the next subsections of the document.

Methodology stage	Steps of policy-making in the context of PolicyCLOUD platform
Problem definition	Step 1: Select the domain and provide a short name to the policy, description of the policy that will be modelled.
Policy discussion	Step 2: Define KPIs for the policy – formula, measure Step 3: Define the stakeholders involved in the policy and other actors
Policy harness	Step 4: Define parameters for KPIs Step 5: Select data source for validation
Policy making	Step 6: Specify analytical tool for the specific scenario (e.g., sentiment analysis) Step 7: Specify the type of graph for the analytical tool Step 8: Define parameters (e.g., region for investigation, time constraints) Step 9: Submit policy (save policy)
Policy insight	Step 10: Run analytical function Step 11: Retrieve the result Step 12: View the heatmap (events are visualized on a map, detailed information for the events is available)
Policy action & impact assessment	Step 13: Evaluate the policy – recall the policy

Knowledge Transformation	-

TABLE 6: MAPPING BETWEEN THE POLICY METHODOLOGY AND THE POLICYCLOUD PLATFORM

4.3.2 The Use Case of the City of Sofia

In this section, Sofia's use case during the PolicyCLOUD project will be used as a practical example for the proposed methodology's validation. In order to present the application of the methodology as clearly as possible, visual examples will be provided, showcasing each methodology stage and the concrete policy-making steps.

The aim of Sofia's use case is to support Sofia Municipality's policy making in important areas of citizen's areas of everyday life by using crowdsourced data, received though Sofia's Municipal contact centre (Call Sofia). By improving policy making in these areas, the overall quality of citizen's life is expected to be improved.

By using the powerful tools provided by PolicyCLOUD platform, Sofia Municipality will be able to carry out detailed analyses, e.g. an analysis of the territorial distribution of the signals by categories /types, areas, districts, major transport roads, etc./, predictive analyses, etc. The results of the analyses allow the municipal and district administrations to identify the problems in the urban environment and adopt or modify adequate policy making-decisions on budget planning and effective use of budget and public resources. They will also help Sofia Municipality focus on improving its policy making related to better control and monitoring in these sectors, as well as preventing/avoiding risky or conflict situations from taking place.

4.3.2.1 Problem Definition

This stage covers Step 1 of the PolicyCLOUD platform, namely Step 1: Select the domain and provide a short name to the policy, and a description of the policy that will be modelled. In the current experiment, the ROAD_INFRASTRUCTURE is the selected domain, where the policy aimed to be created is related to road incidents. Figure 7 shows how the corresponding fields in the PolicyCLOUD platform are initialized at the problem definition stage.

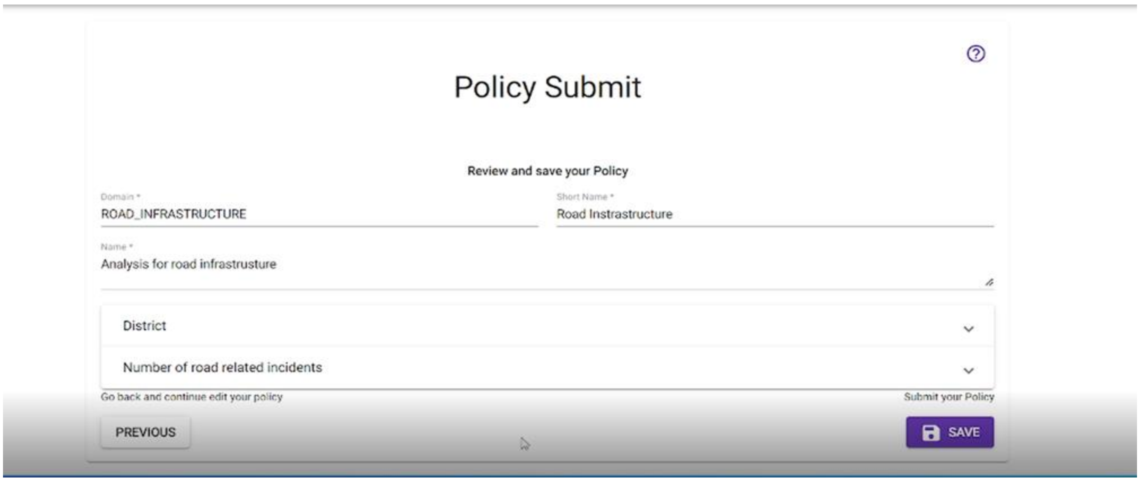


FIGURE 7: PROBLEM DEFINITION THROUGH THE POLICYCLOUD PLATFORM

4.3.2.2 Policy Discussion

This stage covers Step 2 and Step 3 of the PolicyCLOUD platform, namely the definition of KPIs for the policy and the definition of the stakeholders involved in the policy and other actors. The KPI defined is related to the number of road-related incidents. The main stakeholder is Sofia Municipality.

Figure 6 shows how the corresponding fields in the PolicyCLOUD platform are initialized at the policy discussion stage.

FIGURE 8: POLICY DISCUSSION THROUGH THE POLICYCLOUD PLATFORM

4.3.2.3 Policy Harness

This stage covers Step 4 and Step 5 of the PolicyCLOUD platform, namely the definition of the parameters for KPIs and the selection of data sources. Figure 9 shows the dataset used for policy making. It contains information for the road incidents divided into several categories for the period from January 2020 to December 2020.

Category	2020-01	2020-02	2020-03	2020-04	2020-05	2020-06	2020-07	2020-08	2020-09	2020-10	2020-11	2020-12
Липсващи или повредени парапети на мостове, пасарелки и др. елементи с голяма височина	1	2	3	3	5	1	6	6	4	4	5	1
Липсващи или повредени тръбнорешетъчни парапети	26	19	7	11	10	30	32	24	22	24	17	19
Липсващи тротоари	42	41	26	23	35	42	43	46	55	48	36	46
Липсващи, повредени или неефективни конструктивни елементи в подлези	13	13	20	9	20	20	26	22	24	30	8	22
Невъзстановена пътна настилка/ тротоар след извършени ремонтни дейности	72	67	43	55	57	76	81	85	104	112	71	53
Некачествено извършвани строително -ремонтни работи на пътя	13	27	14	25	53	58	91	75	71	72	65	30
Непавомерно поставени кабели по												

Items per page: 20 1 – 13 of 13

FIGURE 9: POLICY HARNESS THROUGH THE POLICYCLOUD PLATFORM (DATASET)

4.3.2.4 Policy Making

This stage covers Step 6, Step 7, Step 8 and Step 9 of the PolicyCLOUD platform, namely specification of the analytical tool for the specific scenario and the type of graph for the analytical tool, definition of parameters required for the analytics and the submission of the policy (save policy)

Figure 10 shows the definition of parameters for the calculation of the corresponding KPI.

Parameters

Analytic Tool Parameter Values:

Name	Type	Value
start_date	EQUAL	2020-01-01 15:23:55.000
end_date	EQUAL	2020-12-31 15:24:12.000
variable1	EQUAL	MONTHYEAR
variable2	EQUAL	TYPE_DESCRIPTION
stacked	EQUAL	True
head	EQUAL	0

FIGURE 10: POLICY MAKING THROUGH THE POLICYCLOUD PLATFORM (PARAMETERS FOR ANALYTICS)

4.3.2.5 Policy Insight

This stage covers Step 10, Step 11 and Step 12 of the PolicyCLOUD platform, namely running the analytical function, obtaining the result and visualisation corresponding graphs. Figure 11 shows how the result is being retrieved. The PolicyCLOUD platform provides a possibility to get an arbitrary analytical result.

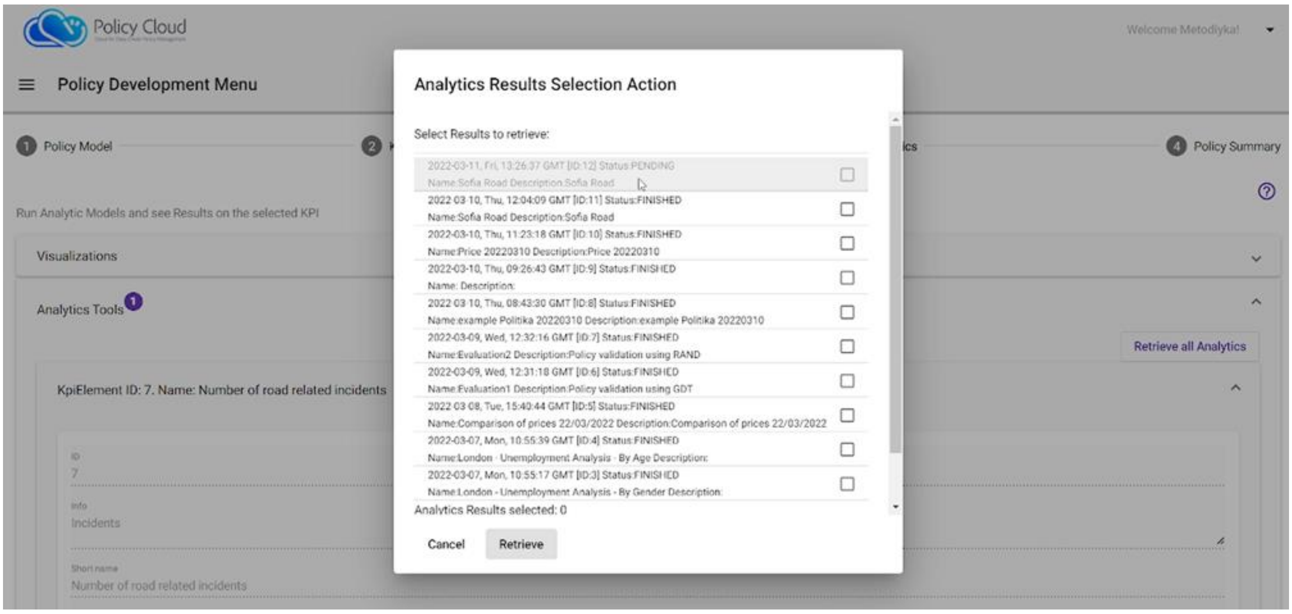


FIGURE 11: POLICY MAKING THROUGH THE POLICYCLOUD PLATFORM (RETRIEVE RESULT)

Figure 12, Figure 13 and Figure 14 below show how insights are obtained through the PolicyCLOUD platform. The frequency of signals, their distribution in time and different categories are presented. All the graphs can be exported together with the corresponding data for further exploration.

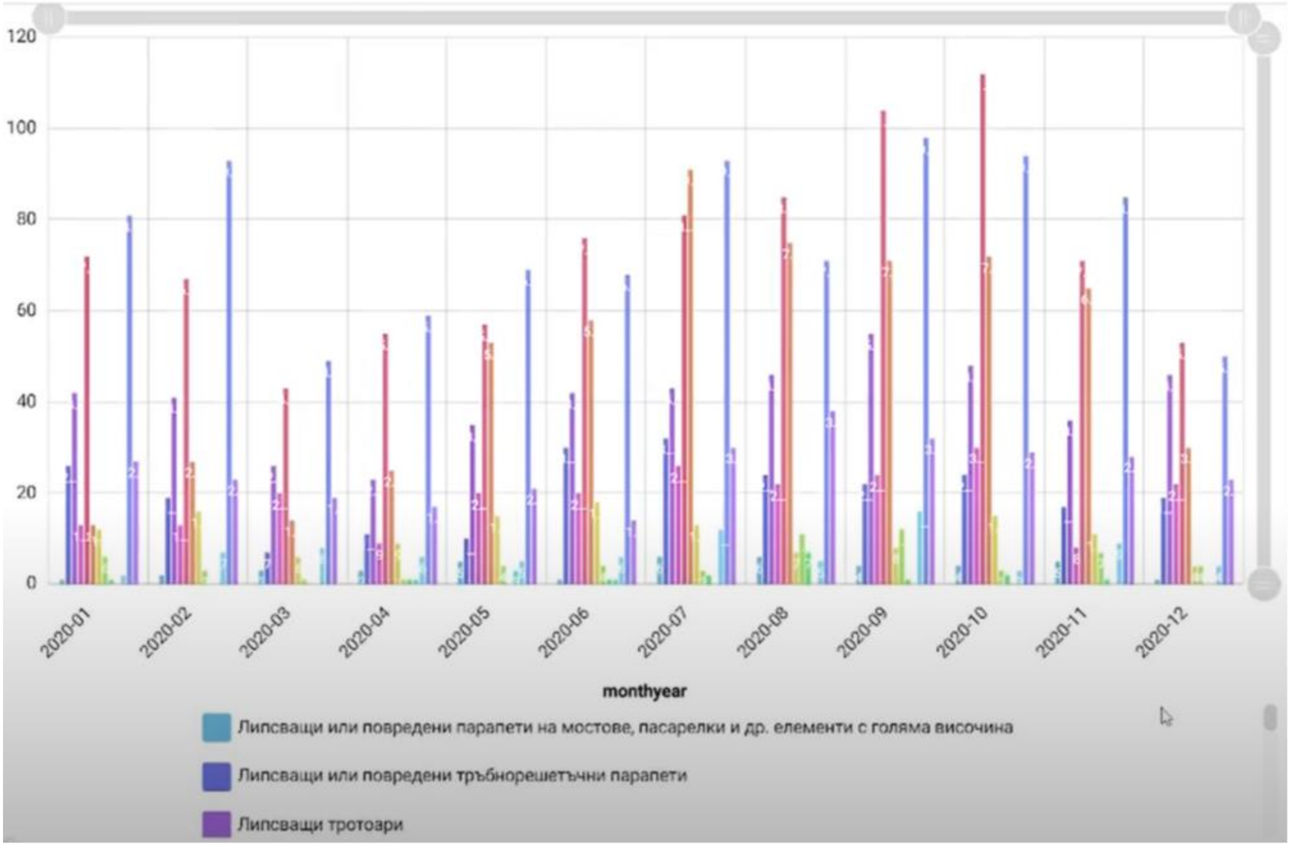


FIGURE 12: POLICY MAKING THROUGH THE POLICYCLOUD PLATFORM (BAR CHART VISUALISATION)

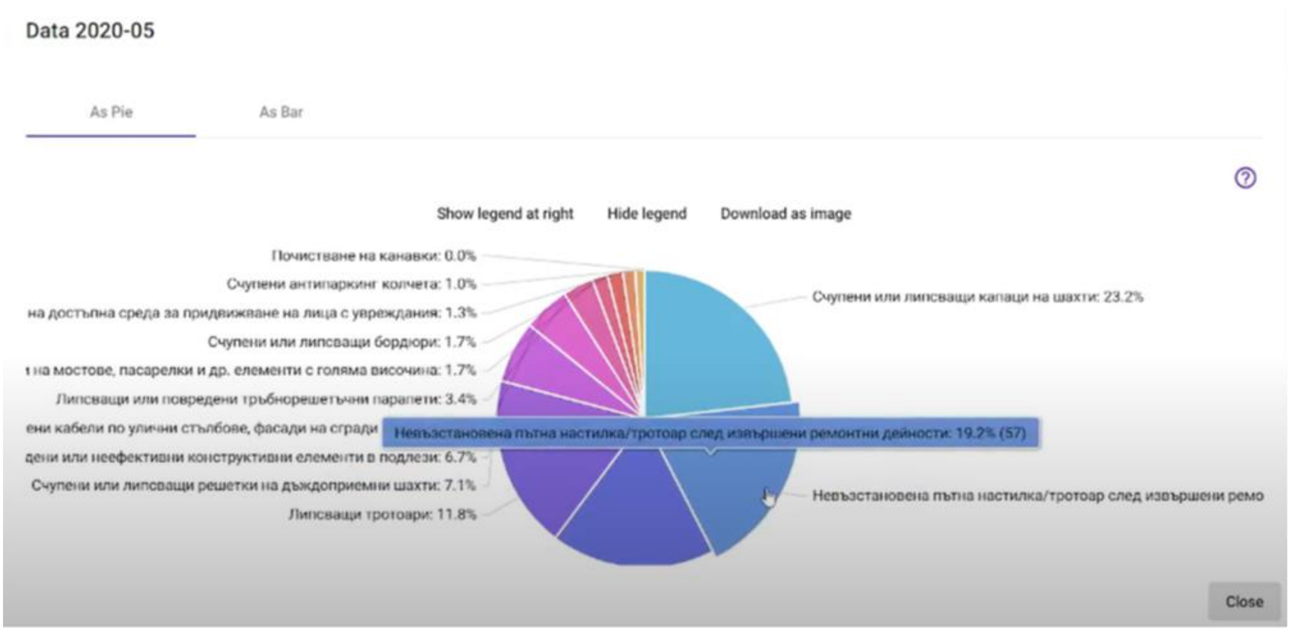


FIGURE 13: POLICY MAKING THROUGH THE POLICYCLOUD PLATFORM (PIE CHART VISUALISAION)

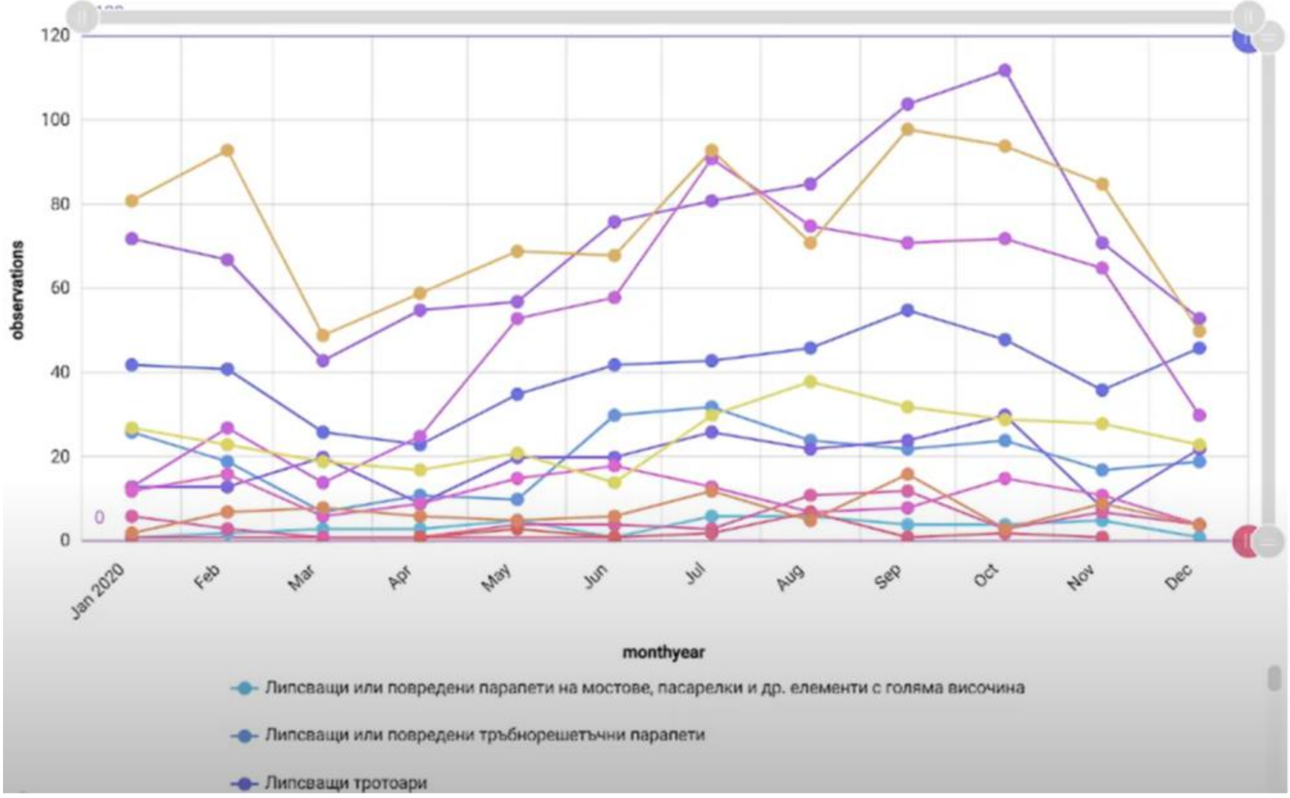


FIGURE 14: POLICY MAKING THROUGH THE POLICYCLOUD PLATFORM (LINE CHART VISUALISATION)

4.3.2.6 Policy Action & Impact Assessment

The analyses produced will support cities in developing and monitoring various policies. The platform allows the continuous tracking of the measures that are applied in a given area of the City of Sofia, as well as the evaluation of their effect.

5 Conclusion

More and more countries, cities, public and private organisations around the world are starting to store and reuse collected data in a cloud environment. The benefits of taking advantage of such digital opportunities are unparalleled and proven to be effective.

Deliverable D 5.8 provided the second iteration of a methodology and a set of adaptable techniques for the creation of policies exploiting Cloud infrastructures and Big Data technologies that was first introduced in Deliverable D 5.1. To this end, the proposed methodology specified how to exploit the value of different mechanisms during the whole policy life cycle in the scope of Cloud environments.

Deliverable D 5.8. followed the structure of Deliverable D 5.1. and was aimed at a wide range of different stakeholders. Thus, apart from the Methodology for policy-making in cloud environments (Section 4), this deliverable included an overview of the main characteristics and benefits of the Cloud and Big Data with an expanded list of concrete examples of good practices and success stories related to their use and utilisation in the public sector in different countries. The provided examples only highlight the importance for the public sector to take advantage of the capabilities and benefits of Cloud technologies and Big Data, and platforms such as PolicyCLOUD in their policy-making

Section 4.3 Methodology Validation was also expanded to include concrete examples from the PolicyCLOUD pilot use case of Sofia of the different stages of the Methodology, their correlation with the steps of the PolicyCLOUD platform, and their practical implementation in regard to policy-making in the context of PolicyCLOUD. Section 4.3 also provided a description of the specific steps required to create a policy model, as well as how the data uploaded to the platform can be analysed using the specific policy model.

References

[1] World Economic Forum (2011), Exploring the Future of Cloud Computing: Riding the Next Wave of Technology-Driven Transformation. https://bit.ly/3fnwOAr, retrieved 2021-06-12

[2] Mell, P., France, T. (2011), *The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology*, NIST Special Publication 800-145. https://bit.ly/3rRwrn0, retrieved 2021-07-12

[3] Cloudflare (2021), *What is the Cloud*? https://bit.ly/3foXrVJ, retrieved 2021-07-12

[4] IBM Cloud Education (2019), *What is FaaS (Function-as-a-Service),* IBM Cloud Learn Hub. https://ibm.co/37hFBzu, retrieved 2021-07-05

[5] Suryavanshi, N. (2017), *What are Cloud Computing Services [IaaS, CaaS, PaaS, FaaS, SaaS],* Medium Official Website. https://bit.ly/3lYRHX3, retrieved 2021-08-09

[6] Red Hat (2021), *What is CaaS?*, Red Hat Official Website. red.ht/3CBOu5x, retrieved 2021-08-09

[7] Iyer, B., Henderson, J. (2010), *Preparing for the future: Understanding the seven capabilities of cloud computing,* MIS Quarterly Executive, Vol. 9 (2), 117-131. https://bit.ly/3BZecjY, retrieved 2021-07-10

[8] Macias, F., Thomas, G. (2011), *Cloud Computing Advantages in the Public Sector: How Today's Government, Education, and Healthcare Organizations Are Benefiting from Cloud Computing Environments*, Cisco White Paper. https://bit.ly/3zYYBz8 , retrieved 2021-07-20

[9] Ostermann, S., (2009), *A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing. CloudComp*. In: Avresky D.R. et al. (eds), (2009), *Cloud Computing. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, 34. Springer, Berlin, Heidelberg. https://bit.ly/2TMHaSS , retrieved 2021-06-03

[10] Müller, S.D., Holm, S.R., Sondergaard, J. (2015), *Benefits of Cloud Computing: Literature Review in a Maturity Model Perspective,* Communications of the Association for Information Systems, 37(1), 851 – 878, DOI:10.17705/1CAIS.03742. https://bit.ly/3A0Pgqt, retrieved 2021-06-14

[11] Berman, S.J., Kesterson-Townes, L., Marshall, A., Srivathsa, R. (2012), *How Cloud Computing Enables Process and Business Model Innovation, Strategy & Leadership,* 40(4), 27-35.

[12] IBM (2009), *The Benefits of Cloud Computing: A new era of responsiveness, effectiveness and efficiency in IT service delivery*. https://ibm.co/3C5FJ3i, retrieved 2021-07-05

[13] IBM Cloud Education (2018), *Benefits of Cloud Computing*, IBM Cloud Learn Hub. https://ibm.co/37dWTO0, retrieved 2021-07-05

[14] Cox, C (2019), *Demystifying the cloud and its benefits for Public Sector organisations,* CTS Blog. https://bit.ly/3rKTFLq, retrieved 2021-07-27

[15] Gartner, Inc.(2021), *IT Glossary - Big Data*. https://gtnr.it/3lqxuZC, retrieved 2021-06-03

[16] Shah, S.I.H., Peristeras, V. & Magnisalis, I. (2021), *DaLiF: a data lifecycle framework for data-driven governments, Journal of Big Data,* 8, 89. https://bit.ly/3ikbuxR, retrieved 2021-07-12

[17] SAS Insights (2021), *Big Data: What it is and why it matters, Big Data Insights.* https://bit.ly/3fpTIaF, retrieved 2021-06-09

[18] Faini, F, Palmirani M. (2016), *Italian Open and Big Data Strategy*, In: Kő A., Francesconi E. (eds) (2016), *Electronic Government and the Information Systems Perspective. EGOVIS 2016. Lecture Notes in Computer Science*, 9831, 105–120. Springer, Cham. bit.ly/3ylq5OR, retrieved 2021-07-13

[19] Munné R. (2016), *Big Data in the Public Sector*, In: Cavanillas J., Curry E., Wahlster W. (eds) (2016), *New Horizons for a Data-Driven Economy*. Springer, Cham. hbit.ly/3yp0sg8, retrieved 2021-07-12

[20] PolicyCLOUD (2021), *Making data-driven policy management a reality across Europe*, PolicyCLOUD Official website. Retrieved from: https://bit.ly/2VwlLOw

[21] PolicyCLOUD (2021), *PolicyCLOUD services*, PolicyCLOUD Official website. Retrieved from: https://bit.ly/3lxKaOu

[22] PolicyCLOUD (2021), *Who benefits?*, PolicyCLOUD Official website. Retrieved from: https://bit.ly/2VwlLOw

[23] European Commission (2020), *Digital Economy and Society Index DESI* (2020), EC Official Website. https://bit.ly/3lpygX5, retrieved 2021-06-24

[24] European Commission (2020), *Digital Economy and Society Index DESI (2020): Italy,* EC Official Website. bit.ly/2VpmlNR, retrieved 2021-06-24

[25] National Interoperability Framework Observatory (2020), *Digital Public Administration Factsheet: Italy* (2020), Joinup Collaborative Platform of the EC . bit.ly/37dG2eb, retrieved 2021-06-25

[26] *Agency for Digital Italy (AgID),* Official Website. https://www.agid.gov.it/, retrieved 2021-06-25

[27] AgID, Department for Digital Transformation of Italy (2020), *Italian Three-Year Plan for IT in the Public Administration 2020-22*, Official Website of the plan. bit.ly/2TUywSy, retrieved 2021-06-25

[28] AgID (2020), *Italian Cloud Strategy of the Public Administration,* AgID Official Website. https://bit.ly/3ylqwbX, retrieved 2021-06-25

[29] Italian Department for Digital Transformation (2020), *Cloud Italia: The Cloud of the Italian Public Administration*, Developers Italia, Official Website. bit.ly/3fpvuNG, retrieved 2021-06-29

[30] Corte dei Conti (2020), *Referto al Parlamento Sullo Stato di Attuazione Del Piano Triennale Per L'informatica 2017-2019 Negli Enti Territoriali della Corte dei Conti*. https://bit.ly/3rKUOmc, retrieved 2021-06-23

[31] European Commission (2020), *Digital Economy and Society Index DESI (2020): Spain,* EC Official Website. bit.ly/3rO0vjn, retrieved 2021-06-24

[32] National Interoperability Framework Observatory (2020), *Digital Public Administration Factsheet: Spain (2020),* Joinup Collaborative Platform of the EC .bit.ly/3rOuUxU, retrieved 2021-06-25

[33] Red.es, Official Website. https://bit.ly/2VidkGH, retrieved 2021-06-25

[34] Regional Government of Aragon (2021), *Consulta Pública previa para elaborar el anteproyecto de ley de medidas para la implantación y desarrollo de la computación en la nube (Cloud Computing) en la Comunidad Autónoma de Aragón.* Official Website of the Regional Government of Aragon. https://bit.ly/37c91iw, retrieved 2021-06-29

[35] Innovation Action COLA (2019), *Inycom Magician meets MiCADO*. https://bit.ly/3ihPbJ6, retrieved 2021-06-29

[36] Inteum (2020), *AST e IECISA despliegan mil escritorios virtuales para el Gobierno de Aragón*. https://bit.ly/3jb7UW1, retrieved 2021-06-29

[37] European Commission (2020), *Digital Economy and Society Index DESI (2020): Bulgaria*, EC Official Website. bit.ly/3igivj8, retrieved 2021-06-24

[38] Bulgarian State E-Government Agency (2021), *About Us,* DAEU Official Website. https://bit.ly/3jhuAUv, retrieved 2021-06-14

[39] National Interoperability Framework Observatory (2020), *Digital Public Administration Factsheet: Bulgaria (2020)*, Joinup Collaborative Platform of the EC. bit.ly/3fmwP7K, retrieved 2021-06-25

[40] Bulgarian State E-Government Agency (2020), *Strategy for the Development of E-government in Bulgaria 2020-25*, DAEU Official Website. https://bit.ly/3ih4L7C, retrieved 2021-06-14

[41] Ministry of Transport, Information Technology and Communications (2019), *Digital Bulgaria 2025 National Programme (2019)*, MTITC Official Website. bit.ly/3liBDyV, retrieved 2021-06-14

[42] Innovative Sofia, Official Website. https://bit.ly/3ygDo3f, retrieved 2021-07-27

[43] Innovative Sofia (2020), *Digital Transformation Strategy for Sofia and Action Plan for Its Implementation*, Innovative Sofia Official Website. https://bit.ly/37fu83G, retrieved 2021-07-27

[44] Innovative Sofia (2021), *Sofia Starts Its Participation in the "Cloud e-Services for the Administration" Project*, Innovative Sofia Official Website. bit.ly/3A1cKvL, retrieved 2021-07-27

[45] Innovative Sofia (2020), *Over 200 Teachers from 16 Schools in Sofia Are Part of the "School in the Cloud" Pilot for Development of Digital Skills and Work in an Online Environment*, Innovative Sofia Official Website. https://bit.ly/3C97hF6, retrieved 2021-07-27

[46] European Commission (2020), *Digital Economy and Society Index DESI (2020): UK*, EC Official Website. bit.ly/3fr3LMn, retrieved 2021-06-24

[47] UK National Cyber Security Centre (NCSC), *What we Do?*, NSCS Official Website. https://bit.ly/3A4dujD, retrieved 2021-07-13

[48] UK National Cyber Security Centre (2018), *Cloud Security Guidance*, NSCS Official Website. https://bit.ly/37lBWR6, retrieved 2021-07-13

[49] UK National Cyber Security Centre (2018), *Implementing the Cloud Security Principles*, NSCS Official Website. https://bit.ly/3ii812N, retrieved 2021-07-13

[50] UK Central Digital and Data Office (2017), *Government Cloud First Policy*, GOV.UK Official Website. https://bit.ly/3A1qsi8, retrieved 2021-07-14

[51] GOV.UK Services (2021), *GOV.UK PaaS: Host your service in the Cloud*, GOV.UK Official Website. https://bit.ly/3ikeCd5, retrieved 2021-07-14

[52] UK Cabinet Office, Government Digital Service (2015), *Digital Marketplace Strategy*, GOV.UK Official Website. https://bit.ly/3rPGTLF, retrieved 2021-07-14

[53] Cabinet Office and Efficiency and Reform Group (2011), *Government Cloud Strategy Policy Paper*, GOV.UK Official Website. https://bit.ly/3zZF5SZ, retrieved 2021-07-14

[54] The Official Portal for European Data, Official Website. bit.ly/3fiFo3u, retrieved 2021-06-03

[55] Open Traffic Global Data Platform, Official Website. opentraffic.io/, retrieved 2021-06-03

[56] Uber Movement Platform, Official Website. https://ubr.to/3A6eR1q, retrieved 2021-06-03

[57] Global Terrorism Database, Official Website. https://bit.ly/2WKcXVH, retrieved 2021-06-03

[58] Italian Presidency of the Council of Ministers, Open Government Partnership (2019), *Fourth National Open Government Action Plan 2019-21*. https://bit.ly/3xgkt7j, retrieved 2021-07-26

[59] Italian Data Protection Authority (2019), *Big Data: Guidelines and Policy Recommendations*. https://bit.ly/3ii9TbJ, retrieved 2021-07-26

[60] National Interoperability Framework Observatory (2017), *Italy facilitates access to public data through portals (#Data4all, SoldiPubblici, ItaliaSicura, OpenExpo),* JoinUp Collaborative Platform of the EC, Official Website. https://bit.ly/3C3JHtg, retrieved 2021-07-26

[61] SoldiPublicci, Official Website. https://bit.ly/3C8rSt9, retrieved 2021-07-26

[62] ItaliaSicura, Official Website. https://bit.ly/2VtVgt3, retrieved 2021-07-26

[63] OpenExpo 2015 Milano, Official Website. http://www.expo2015.org/, retrieved 2021-07-26

[64] Italian National Open Data Portal, Official Website. https://dati.gov.it/, retrieved 2021-07-26

[65] Km4City, Official Website. https://www.km4city.org/, retrieved 2021-07-26

[66] Open Data Lombardy, Official Website. https://dati.lombardia.it/, retrieved 2021-07-26

[67] Spanish Ministry of Economic Affairs and Digital Transformation (2020), *2025 Digital Spain Strategy,* Official Website of the Spanish Ministry of Economic Affairs and Digital Transformation. https://bit.ly/3rPaSn5, retrieved 2021-07-27

[68] Spanish Ministry of Territorial Policy and Public Function (2020*), Fourth Action Plan of Spain of the Open Government Partnership 2020-24*. https://bit.ly/3C6TkaF , retrieved 2021-07-27

[69] Joinup Collaborative Platform of the European Commission (2021), *About Aragon Open Data Pool: Transform Open Data to Linked Open Data and Browse it!*, Joinup Official Website. https://bit.ly/3xnECIG, retrieved 2021-07-27

[70] Datos (2020), *How to implement CKAN: Real Case of the Aragon Open Data portal*. https://bit.ly/3yqhmLG , retrieved 2021-07-27

[71] The Open Data Initiative of the Government of Spain, Official Website. www.datos.gob.es, retrieved 2021-07-27

[72] The Official Portal for European Data (2019), *Badajoz: a digital success story from Spain.*, Official Website of the Portal for European Data. https://bit.ly/3C7iANX, retrieved 2021-07-27

[73] Bulgarian Open Data Portal (2021), *Accessibility*, Bulgarian Data Portal Official Website. https://bit.ly/3fq0ddz, retrieved 2021-07-09

[74] Bulgarian Ministry of Transport, Information Technology and Communications (2019), *Digital Transformation of Bulgaria 2020 - 2030 Strategy,* Official Website of MTITC. https://bit.ly/3ii8zpn, retrieved 2021-07-09

[75] Bulgarian Council of Ministers (2020), *"Bulgaria 2030" National Development Programme*, National Public Consultations Portal. https://bit.ly/3ikxatq , retrieved 2021-07-09

[76] Bulgarian Open Data Portal, Official Website. https://data.egov.bg/, retrieved 2021-07-09

[77] Innovative Sofia (2021), *Sandbox for Innovative Solutions: Highlights, Useful Links, and Answers,* Innovative Sofia Official Website. https://bit.ly/3iefzU5, retrieved 2021-07-09

[78] UK Department for Digital, Culture, Media & Sport (2020), *UK National Data Strategy*, GOV.UK Official Website. https://bit.ly/3fiTkKH, retrieved 2021-07-12

[79] *UK Official Open Data Portal*, Official Website. https://data.gov.uk/, retrieved 2021-07-12

[80] National Interoperability Framework Observatory (2020), *Digital Public Administration Factsheet: UK (2020),* Joinup Collaborative Platform of the EC. bit.ly/37bLyOB, retrieved 2021-06-25

[81] UK National Archives (2021), *UK Government Licensing Framework,* UK National Archives Official Website. https://bit.ly/3zSCOcq, retrieved 2021-07-14

[82] UK National Archives (2021), *Open Government License for Public Sector Information*, UK National Archives, Official Website. https://bit.ly/3AffSEz, retrieved 2021-07-14

[83] Open Data Camden, Official Website. https://bit.ly/ 3xij3Jv, retrieved 2021-07-14

[84] Curry E. (2016), *The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches*, In: Cavanillas J., Curry E., Wahlster W. (eds), *New Horizons for a Data-Driven Economy*. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3, retrieved 2021-06-01

[85] Faroukhi, A.Z., El Alaoui, I., Gahi, Y. et al. (2020), *Big Data monetization throughout Big Data Value Chain: a comprehensive review,* J Big Data 7 (3). Springer, Cham. https://doi.org/10.1186/s40537-019-0281-5, retrieved 2021-07-05

[86] Koo, J., Kang, G., Kim, Y.G. (2020), *Security and Privacy in Big Data Life Cycle: A Survey and Open Challenges. Sustainability,* 12 (24). https://doi.org/10.3390/su122410571, retrieved 2021-07-19

[87] Big Data Public Private Forum (2017), *D2.2.2 Final Version of Technical White Paper*. https://bit.ly/3jltjMl, retrieved 2021-06-23

[88] PolicyCLOUD. *D4.3 Reusable Model & Analytical Tools: Design and Open Specification-2.* Moatti Yosef. 2021.

[89] Policy Cloud. *D3.6 PolicyCLOUD's Societal and Ethical Requirements & Guidelines*. Audino Alice. 2021.

[90] Siddiqa, A. et al. (2016), *A survey of Big Data management: Taxonomy and state-of-the-art,* In: Journal of Network and Computer Applications, Vol. 71, pp. 151-166.

[91] SKOS Simple Knowledge Organization System - Reference (W3C, 2009). https://bit.ly/3ynzVzN, retrieved 2021-06-29

[92] Jaulent M., D. et. al. (2018), *Semantic interoperability challenges to process large amount of data perspectives in forensic and legal medicine*, Journal of Forensic and Legal Medicine, 57. https://doi.org/10.1016/j.jflm.2016.11.002, retrieved 2021-06-23

[93] Smart Data Models Initiative, FIWARE Official Website. bit.ly/2VgXbRV, retrieved 2021-07-05

[94] CityGML Standard, OGC Official Website. https://bit.ly/3lrzJfd, retrieved 2021-07-05

[95] Venkatram K., Geetha. M. (2017), *Review on Big Data & Analytics – Concepts, Philosophy, Process and Applications*, In: Cybernetics and Information Technologies, 17 (2), 3-27

[96] Fang, R., Pouyanfar, S., Yang, Y., Chen, S., Iyengar, S.S. (2016), *Computational Health Informatics in the Big Data Age: A Survey*, In: ACM Computing Surveys, 49 (1)

[97] Hordri N. F., Samar, A., Yuhaniz S.S., Shamsuddin S.M. (2017), *A Systematic Literature Review on Features of Deep Learning in Big Data Analytics*, In: Int. Journal of Advanced Soft Computing Applications, 9 (1), 32-49. ISSN 2074-8523

[98] Irfan, S., Babu V.B. (2016), *Information retrieval in Big Data using evolutionary computation: A survey,* In: Proceeding of IEEE International Conference on Computing, Communication and Automation, ICCCA, 208-213.

[99] Hartwig, F., Dearing, B.E. (1979), *Exploratory Data Analysis*, SAGE, p.83.

[100] Pawar A.B., Jawale M.A., Kyatanavar D.N. (2016), *Fundamentals of Sentiment Analysis: Concepts and Methodology*, In: Pedrycz W., Chen S.M. (eds), *Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence*, 639. Springer, Cham. https://doi.org/10.1007/978-3-319-30319-2_2

[101] Mastoli, M. M., Pol, U.R., Patil, R.D. (2013), *Machine Learning Classification Algorithms for Predictive Analysis in Healthcare*, International Research Journal of Engineering and Technology, 6 (12). https://bit.ly/3fpYGE4p , retrieved 2021-06-25

[102] MacGregor, J. (2013), *Predictive Analysis with SAP®p* Bonn: Galileo Press.

[103] Harvey, A. (2014), *Trend analysis*, In: Wiley StatsRef: Statistics Reference Online, 1-21

[104] Budak, C., Agrawal, D., El Abbadi, A. (2011), *Structural trend analysis for online social networks*. Proceedings of the VLDB Endowment, 4(10), 646–656. https://doi.org/10.14778/2021017, retrieved 2021-06-25

[105] Caldarolae., A., Rinaldi, M. (2017), *Big Data Visualization Tools: A Survey: The New Paradigms, Methodologies and Tools for Large Data Sets Visualization*, In: Proceedings of the 6th International Conference on Data Science, Technology and Applications, 296-305. ISBN: 978-989-758-255-4

[106] Zhang L. et al. (2012), *Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems,* 2012 IEEE Conference on Visual Analytics Science and Technology, 173-182. ISBN: 978-1-4673-4753-2

[107] PolicyCLOUD*. D5.2: Cross-Sector Policy Lifecycle Management: Design and Open Specification 1.* Duzha Armend. 2020

[108] Policy Cloud*. D5.4: Cross-Sector Policy Lifecycle Management: Design and Open Specification 2.* Duzha Armend. 2021

[109] European Commission (2019), *High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI*, EC Official Website. https://bit.ly/37q0NTT, retrieved 2021-06-28

[110] European Union Agency for Cybersecurity-ENISA (2020), *EUCS-Cloud Services Scheme*, ENISA Official Website. https://bit.ly/3A6C4R3, retrieved 2021-06-28

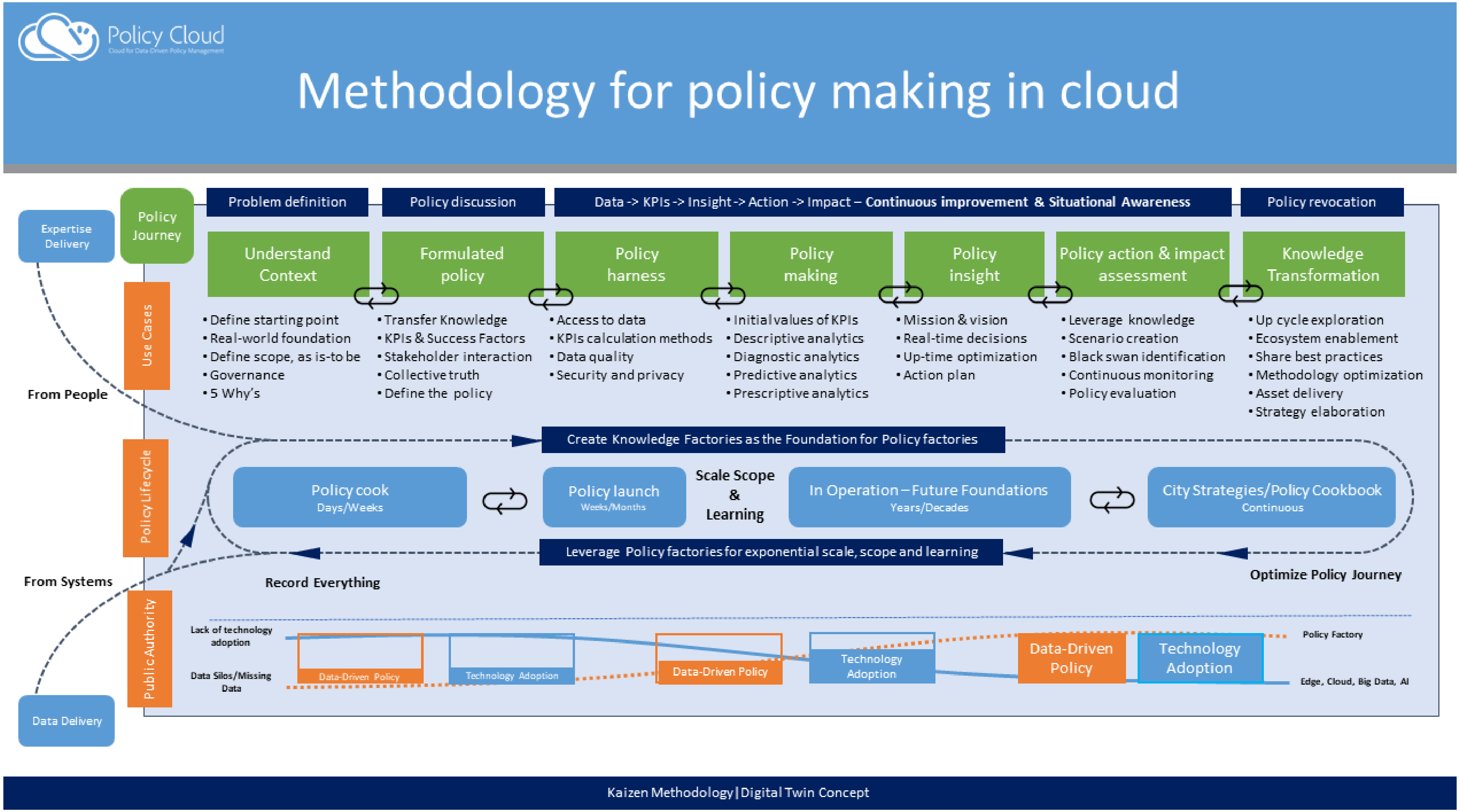
[111] Karta Legal (2021), *What is Kaizen Methodology?*, Karta Legal Official Website, https://bit.ly/37iaHHq, retrieved 2021-07-01

[112] DUET Digital Twins (2021), *What is a Digital Twin?*, DUET Official Website. https://bit.ly/3yzgq7P, retrieved 2021-07-01

[113] International Data Spaces Association, https://internationaldataspaces.org/

[114] Gaia-X Initiative (2022), Official Website. https://gaia-x.eu/what-is-gaia-x/core-elements/data-spaces/, retrived 2022-07-22

Appendix A



APPENDIX A: POLICYCLOUD METHODOLOGY FOR POLICY MAKING IN CLOUD ENVIRONMENTS