SPECIAL

Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance

Deliverable 1.4

Technical Requirements V1

Document version: 1.0

SPECIAL DELIVERABLE

Name, title and organisation of the scientific representative of the project's coordinator:
Mr Philippe Rohou t: +33 4 97 15 53 06 f: +33 4 92 38 78 22 e: philippe.rohou@ercim.eu

GEIE ERCIM, 2004, route des Lucioles, Sophia Antipolis, 06410 Biot, France
Project website address: http://www.specialprivacy.eu/

Project	
Grant Agreement number	731601
Project acronym:	SPECIAL
Project title:	Scalable Policy-awarE Linked Data arChitecture for prIvacy, trAnsparency and compLiance
Funding Scheme:	Research & Innovation Action (RIA)
Date of latest version of DoW against which the assessment will be made:	17/10/2016
Document	
Period covered:	M1-M8
Deliverable number:	D1.4
Deliverable title	Technical Requirements V1
Contractual Date of Delivery:	31/08/2017
Actual Date of Delivery:	01/09/2017
Editor (s):	Bert Bert Van Nuffelen
Author (s):	Bert Bert Van Nuffelen
Reviewer (s):	Piero Bonatti, Sabrina Kirrane
Participant(s):	Bert Bert Van Nuffelen, Piero Bonatti, Sabrina Kirrane
Work package no.:	1
Work package title:	Use Cases and Requirements
Work package leader:	CeRICT
Distribution:	PU
Version/Revision:	1.0
Draft/Final:	Final
Total number of pages (including cover):	29

Disclaimer

This document contains description of the SPECIAL project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the SPECIAL consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

SPECIAL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731601.

Table of Contents

1 Introduction

This document reports on the technical requirements and challenges for the SPECIAL platform. Building upon:

- Deliverable 1.1 describing the use-cases,
- Deliverable 1.2 describing the legal context and the initial legal analysis of the use-cases, and
- Deliverable 3.1 on the first release of the SPECIAL platform

this deliverable presents an overarching technical perspective of the SPECIAL platform. An initial version of this perspective can be found in Deliverable 3.1, but in this report it is further elaborated and placed in a broader context, called the SPECIAL ecosystem. Key stakeholders are identified and their main interactions with the platform are described as user stories. Additionally, we detail the to-be applied approach for privacy threat assessment and the to-be applied risk mitigation strategies.

The objective of this deliverable is to facilitate the upcoming development and research work by the consortium. The deliverable forms a pair with Deliverable 1.3. Deliverable 1.3 provides a state-of-the art analysis on consent management, policy language and transparency. This deliverable elaborates the software architectural perspective. Together they will lead to the creation of a development roadmap for the project. Considering the iterative nature of the project, this deliverable is not meant to serve as a complete list of requirements, but rather as a summary of our initial analysis, that will be updated regularly as the project advances. The extended document will subsequently be published as D1.8 Technical requirements V2 at the end of month seventeen.

2 Use Cases and the SPECIAL approach

The SPECIAL project is motivated from the need for a simplified personal data management that complies with the General Data Protection Regulation (GDPR). Three use-cases partners, two from the telco industry: Proximus and TLabs, and one active in financial data services industry: Thomson Reuters Limited, have described their ideas on value adding services in D1.1 *Use Case Scenarios*.[1] These business objectives can only be realised if the personal data required to fuel the services is properly managed.

By the GDPR, businesses are granted the ability to create added value from data including data of a personal nature provided by the data subjects (from whom personal data is collected and processed) are given control on their own personal data. Among others, the GDPR states that control on the usage of personal data implies that the purposes for which the data are being acquired are understandable, permission to use the data is obtained in a comprehensive way and that the actual usage is verifiable. For SPECIAL, the control takes the form of **consent** and policy data management to capture the data subject's permissions for personal data processing and sharing, data management of the data usage traces to provide **transparency** on the usage and **compliance** mechanisms to guarantee and prove that the usage is in accordance with the given permissions and the legislation.

Based on the use-case descriptions and the additional provided background information, this deliverable presents an overarching SPECIAL data processing ecosystem with integrated support for consent, transparency and compliance. Each use case corresponds to an instance reusing common parts but differentiating on the used data and the service being implemented.

SPECIAL technical objective is to realise consent, transparency and compliance mechanisms for big data processing. Therefore, the service aspect defined by the use case instances will only be implemented to the level they can be used to demonstrate the consent, transparency and compliance mechanisms.

To obtain the desired personal data management and processing, SPECIAL defines an approach based on policy aware data processing, which is shown in the figure below.

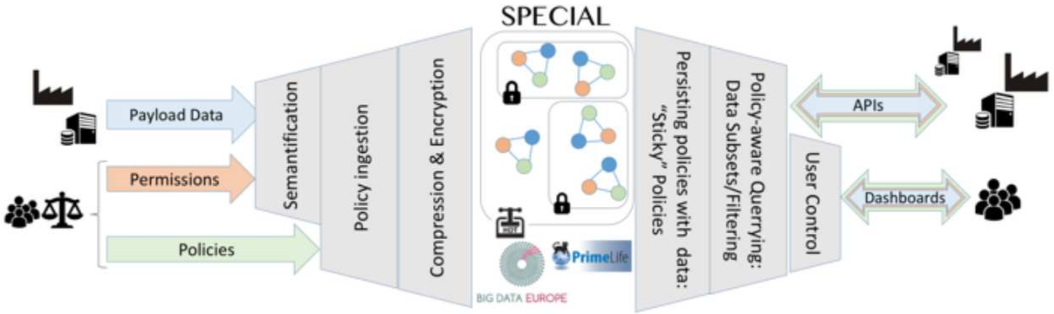


Figure 1:Birds-eye View of the SPECIAL Approach

This approach, from left to right, is defined in short as follows:

- first, harmonise the data (both payload and the consent data) by making the semantics of the data explicit,

[1] Due to confidenciality no details of the individual usecases are presented.

- then, augment the data with consent approval and usage policies,
- ensure that the data is securely and efficiently accessible (by applying techniques such a compression, encryption etc.),
- this creates data with sticky policies[2], and
- finally, provide Application Programming Interface (API's) and User Interfaces (UI's) to access the payload according to the associated consent and applicable policy.

Using the proposed approach, the payload data processing is integrated with the consent and policy data. While control is awarded to the data subject via transparency and compliance checking mechanisms. If implemented correctly the system has a by-design guarantee that the data subjects consent is honoured.

[2] Sticky policies is the term for the approach to attach the policy to the data in a manner that ensures that the policy is tightly coupled to the data (which is especially important when data transcends company boundaries).

3 SPECIAL Ecosystem

The GDPR defines a data processing ecosystem consisting of various stakeholders (such as, data subjects, data controllers and data processors, supervisory authorities), and legal rights, obligations and constraints with respect to the personal data processing. The SPECIAL ecosystem is an instance of the GPDR data processing ecosystem using the data subject’s consent and using the data processor’s transparency information, provided by the data controller/processor, to verify compliance with the legislation.

Thus, central in SPECIAL is the management of consent and transparency data. This area is responsible for recording and managing the data subject’s consent, administering the policy definitions, providing data for audits, supporting the compliance verification, etc. From now on we will use the abbreviation CTC, referring to Consent-Transparency-Compliance, to denote the area of work to which the SPECIAL project is devoted.

The other area of work in the SPECIAL ecosystem is the data processing which takes the consent into account. Whereas the CTC management is mostly domain neutral and common for each of the use cases, the added value service data processing is specific to the business objectives. In this area of work, SPECIAL will provide a common methodology and several libraries that are required to enable the implementation of the different services use cases. This area will be referred from now on to as the AV, the business Added-Value data processing.

In the following we will further elaborate on the SPECIAL ecosystem. The next sections detail the functional/technical requirements more concretely.

3.1.1 Stakeholders

In the context of SPECIAL, the GDPR defines the following key stakeholders:

- Data Subject: the person who is sharing personal data
- Data Controller: the organisation which owns the data processing service
- Data Processor: the organisation which actually processes/stores the data. This may be different from the Data Controller, e.g. a cloud service provider.
- Supervising Authority: the authority responsible for auditing if the data processing happens according to the legislation.

Without restricting the validity of the SPECIAL outcomes, we can simplify the SPECIAL ecosystem by assuming that the data controller and the data processor are the same entity. In the following, the term *service offering company* is therefore used as alternative term for the data processor or data controller.

Next we will detail the data subject and the data controller in two distinct roles. The data subject can have the role of a *personal data provider* or as *data service consumer*. Indeed, one can be a personal data provider without consuming the result. For instance, you may grant your telco operator permission to use your communication data for the creation of a traffic pattern knowledgebase, but not consume the service exploiting that traffic pattern. On the other hand, a service consumer might not be a personal data provider. For example, when traffic data is used as the basis to create announcements emitted as radio messages. Obviously within SPECIAL the first role is the most critical one and will be denoted as the data subject.

The data controller is divided in two roles: one is the *policy administrator* and the other is the *service developer*. The policy administrator has the responsibility to maintain and enforce the policies that are associated with the to-be gathered personal data. This is a key role as the policy administrator will translate both the business objectives and the legation obligations into a machine processable format. From them, SPECIAL will design the base ontological framework capturing all necessary aspects of for CTC data management compliant with the GDPR. Service developers are responsible for the service implementation. They expect to find within the SPECIAL ecosystem libraries, APIs and guidelines which can be used to build GDPR compliant AV-services.

3.1.2 SPECIAL ecosystem

Figure 2: SPECIAL Ecosystem depicts the SPECIAL ecosystem. The left, coloured in yellow, is the CTC management area. The right side, in blue, is the AV data processing area. They are connected via secure interfaces via which CTC data is exchanged.

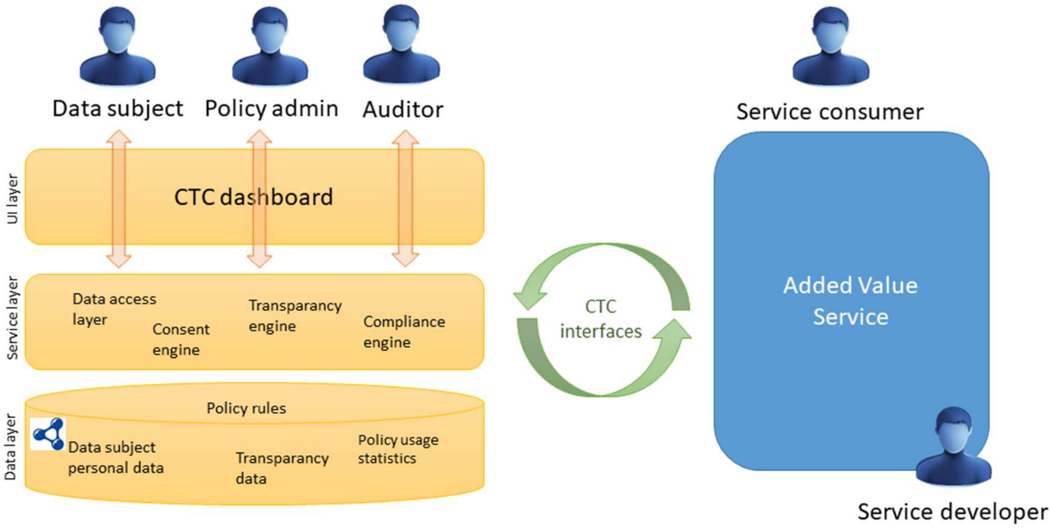


Figure 2: SPECIAL Ecosystem

The figure shows the interaction of the identified roles within the SPECIAL ecosystem. Three roles interact with the CTC dashboard:

- Data subject
- Policy admin
- Auditor

Via the CTC dashboard the data subject can execute the control on the usage of its personal data. Consent can be given or withdrawn, the purpose (policy context) for which consent is requested can be explored, insight to the usage of the data is given, etc. The policy admin is given the power to manage the policies and the power to verify the compliance of the AV service to the policy definitions. For the Auditor, the CTC dashboard provides the necessary verification to ensure compliance with the legislation.

The system architecture for the CTC management follows the multi-tier pattern. In the following subsection more details of each layer is given. In short, from top to bottom: the UI layer implements the UI interaction for the different user roles; the service layer provides the services for accessing data, consent and policy management, transparency and compliance verification; the data layer is responsible for storing the data securely.

The AV data processing area contains the data processing system, developed by the company’s service developer, creating the added value data for the company. The service consumer is the party who consumes the AV service. More detail about our vision for the AV data processing is found in subsection **Error! Reference source not found.**.

The AV data processing will implement the SPECIAL approach for policy aware data processing. For that it must interact with CTC management via CTC service layer. This interaction is denoted in green. Initial thoughts on the interaction been company systems and the SPECIAL components are presented in Deliverable D1.3.

3.1.3 Linked Data centric

The above introduced SPECIAL ecosystem is from a birds-eye perspective comparable to other approaches. It distinguishes from others by the application of Linked Data[3] (or Semantic Web) as the technical foundation.

The following benefits from Linked Data form the basis for our decision to use it:

- it is based upon a domain neutral, flexible, multi-lingual data representation format standardised by W3C,
- it is the most popular data ecosystem supported with automated reasoning capabilities (OWL[4]) that has been standardised[5],
- it well-balances the human readable aspect with machine readable aspect,
- it is web-enabled by design,
- it is ideal for data integration tasks, and
- it is well-suited for cross-system/cross-organisational data interoperability.

The last item has a not to-be under-estimated value for community adoption. Since the personal data, the consent to use it and the associated policy is going to be used by many different systems within the service offering company, but also across company borders a common, reliable, semantically unambiguous way to reference this data is an important requirement. Otherwise desired properties such as transparency, which requires data processing and sharing events to be associated with the corresponding consent, are hard to achieve.

Consequently, this design requirement influences the component design for SPECIAL ecosystem. In particular, that data processing technology which does not natively support Linked Data has to be extended with it. In the SPECIAL approach this is called Semantic Lifting. Vice versa, Linked Data native components[6] might not have the required data privacy or data security properties. It might be necessary to extend those components before they can be part of the SPECIAL ecosystem.

3.2 Consent, Transparency and Compliance Management

The focal point of the SPECIAL project is the consent, transparency and compliance management. *Figure **3**: CTC dashboards* highlights this area in more detail.

[3] https://www.w3.org/standards/semanticweb/data, the term Linked Data and Semantic Web are used here as synonyms.

[4] https://www.w3.org/TR/owl2-primer/

[5] To our knowledge the only one.

[6] With components we refer to all aspects: from vocabularies, standards, protocols to software implementations.

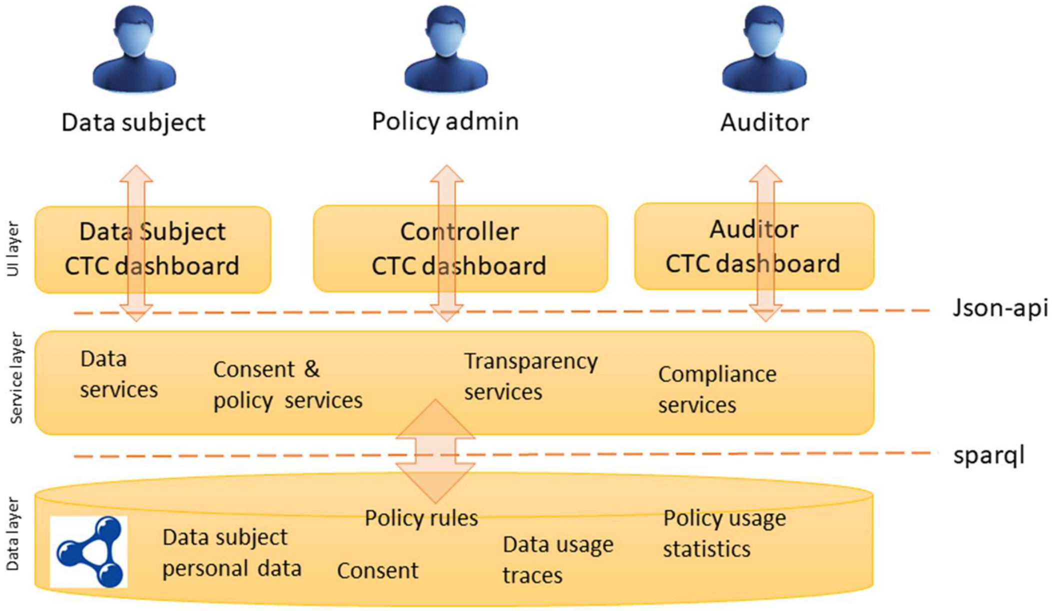


Figure 3: CTC dashboards

Data layer

The *data layer* manages and stores the CTC data which covers among others policy rules (usage constrains, legislative obligations and constraints, business logic), personal data of the data subject, and his/her consent for the data use, provenance trails about the data processing for transparency, etc. The data layer will be based on Semantic Web technology (RDF, OWL). We refer to Deliverable 1.3 for a first, deeper analysis of the data layer covering functional and technical requirements, the concepts to be captured, an overview of existing approaches, the challenges and implementation considerations.

Service layer

The *service layer* is responsible for facilitating the creation and the access to the CTC data. The base functionality are interfaces assisting the implementing of the UI. More advanced services support the consent interpretation, transparency insights and compliance verification. *Table 1 advanced CTC services* gives an overview of the advanced services we foresee to be implemented. The service layer is also the bridge between the Linked Data based data-layer and the other data representations commonly used in the practice. For instance, the de facto standard for data exchange in UI implementation frameworks is json. More variety is expected in the implementation of the added-value services. For the interface two design principles are applied: (a) whenever possible a standard is applied and (b) preference goes to already used standards in the SPECIAL ecosystem. *Figure 3: CTC dashboards* shows that the interface between the service layer and the data layer is SPARQL[7] and between the service layer and the UI layer json-api[8]. SPARQL is a natural choice for interfacing with Linked Data. Json-api is an industry standard driven by the Ember framework community[9].

[7] https://www.w3.org/TR/sparql11-query/

[8] http://jsonapi.org/

[9] https://emberjs.com/

Component	functionalities
Transparency engine	• List the data processing and sharing events happened • Find data processing and sharing events by data subject, by consent, by temporal window • Add data processing and data sharing events to the transparency ledger • Export the transparency data in an interoperable format
Consent engine	• List the data subject's consent timeline (when given consent, when retracted, etc.) • Fold/unfold consent into/from groups • Register consent • Revoke consent • Get all contextual information about a consent to create a human readable view • Associated a data processing event with the consent
Compliance engine	• Coherency validation of transparency data and consent data • Get statistics for key parameters (#consents, #revocations, #data sharing events, #data processing events …)

Table 1 advanced CTC services

UI layer

The top layer in *Figure 3: CTC dashboard* is the *UI-layer*. We foresee independent UI's serving the needs for each role. This simplifies the overall access-control mechanism as the interface targets only a single kind of users. Additionally it creates a separation of concerns reducing the risk of disclosing information. The table below lists the most important interactions of each role within the system's policy management dashboard. Section 5 gives a more elaborated analysis on (motivations for) key interactions for the data subject's CTC dashboard.

Role		
	explore	**change**
Data Subject	• browse personal data • browse transparency data • explore the policy definitions • notifications about change & requests status	• adapt policy choices • request to change key personal data • request data transparency • request to be forgotten
Controller Admin	• see impact of policy changes	• define policies
Auditor	• see statistics per policy • see processes related to the execution of the policy	• none

Table 2: SPECIAL Ecosystem Interactions

3.3 Added Value Service

A second area of the SPECIAL ecosystem is the AV service for which the data subject's consent has been gathered.

SPECIAL enables the creating of privacy preserving added-value services, that enables data to be combined, aggregated, analysed, etc. The origin of the data may be very diverse: ranging from public open accessible data (e.g. touristic activity statistics from the national statistical office), commercially acquired data (e.g. events happening in a region), to data obtained from other services owned by the company (e.g. location data from the telco network). In order to enable companies to informed privacy preferences and legal obligations, the data needs to be connected and combined with both consent (obtained from data subjects) and policy rules (derived from usage constraints and legal obligations) that state how the data can be used.

The use cases described by our use case partners (see Deliverable D1.1) show a wide diversity of services that could leverage our SPECIAL ecosystem. Independent of the privacy aspects the data processing must address several big data challenges because of the characteristics of the data itself. These data characteristics are commonly called the four Vs of Big Data:

- *Volume*: the amount of data being processed,
- *Velocity*: the speed that data is provided,
- *Variety*: the different models/formats in which the data is provided
- and *Veracity*: the trustworthiness of data.

Concerning volume and velocity, the data processor must handle large amounts of data, as the use cases indicate constant data streams in great amounts. Streaming processing support is hence required. But at the same time support is required for processing less voluminous, yet complex data having a low change rate.

All use cases indicate the usage of several data sources provided by as many different systems. To address the heterogeneity of the data sources, semantic web technologies will be applied too. This creates a uniform data layer easing the interaction with the policy management data.

In terms of veracity, some use cases provide data that is readily available and easily understood as the data is under the control of the use case partner. However, data may also be collected from "open, uncertain sources". In that case the quality and trustworthiness of the data must be investigated before they can be integrated in the service.

3.3.1 The Lambda Architecture

Within the Big Data community, the lambda architecture[10] is a standard architecture pattern for handling large quantities of data. The lambda architecture, which is depicted in Figure 4, is a term given by Nathan Marz for a generic, scalable and fault-tolerant data processing architecture, based on his experience working on distributed data processing systems at Twitter. It distinguishes between three layers: the serving layer, the batch layer and the speed layer.

[10] http://lambda-architecture.net/

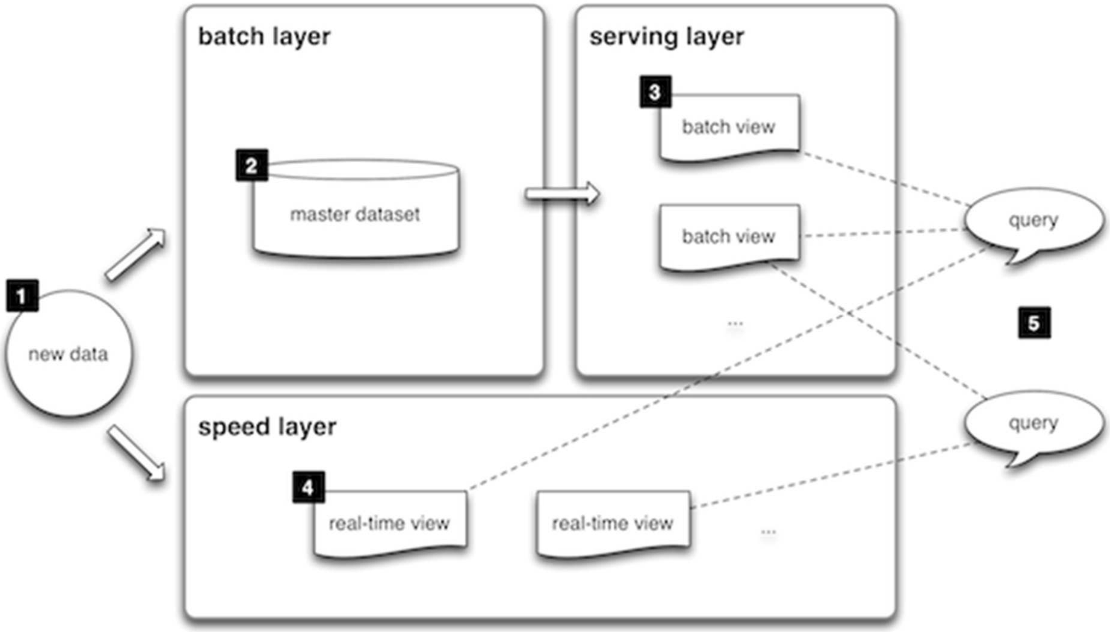


Figure 4 the Lambda Architecture pattern (as defined by Nathan Marz)

The new data is processed by the batch layer and speed layer to realise derived data views. The serving layer is responsible to make the views efficiently queryable for the business applications. The batch layer and speed layer perform data processing, but where-as the batch layer is optimised for performing processing on high amounts of data at once renewed with a low frequency, the speed layer is optimised for performing processing small amount of data given in a high frequency.

Tasks within the batch layer normally require a substantial amount of time to finish. The resulting data view can be a final product to be used in the service layer, but often and it is expected to happen in SPECIAL, it also acts a preprocessing step for the speed layer. Then it lays out the data so that the speed layer (optimised to handle a high volume of messages having a small data payload) can work efficiently.

3.3.2 SPECIAL Lambda Architecture

Within SPECIAL, the serving layer will be simplified to deliver the data views on which the desired customer facing service can be built. SPECIAL efforts will focus on supporting the batch and speed layer to enable data processors to integrated with policy enforcement and compliance components. That means integrating support for associating policies with the payload data, integrating policy enforcement and compliance checking mechanisms.

To integrate with our SPECIAL CTC management, semantic lifting is required. This means the lambda architecture will be augmented with Linked Data processing capabilities[11].

The diagram below shows the SPECIAL lambda architecture highlighting the batch layer. The speed layer and serving layer are greyed. The service background information represents a collection of data sources required for the batch processing. First, data is semantically lifted by transforming the data into RDF format, and then aggregated with policy data and background

[11] In Deliverable 3.1, the Semantic Data Lake Ontario has been discussed. Some aspects of this might be applicable here too, but that has to be investigated.

data. Finally, data is partitioned to organise it for optimal access and future processing. To simplify the diagram, the resulting data is offered only to the speed layer processing.

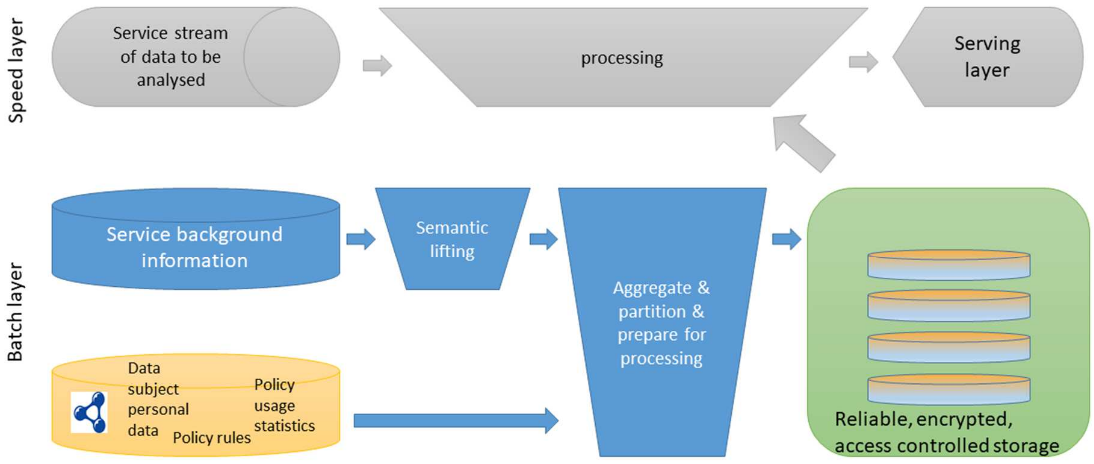

Speed layer
Service stream of data to be analysed
processing
Serving layer
Batch layer
Service background information
Semantic lifting
Aggregate & partition & prepare for processing
Data subject personal data
Policy rules
Policy usage statistics
Reliable, encrypted, access controlled storage

Figure 5: SPECIAL lambda architecture

The speed layer provides streamed data processing, which is done through a pipeline. A stream of data consists of small pieces of data, called a batch[12], that are processed through the data processing pipeline. A prototypical pipeline for SPECIAL, shown in *Figure **6**: SPECIAL Streaming Data Processing*, will contain the following sequence:

1. First, apply semantic lifting on each data batch in the stream;
2. Second, enrich the data batch with the necessary background information;
3. Finally, calculate the desired result for each data batch.

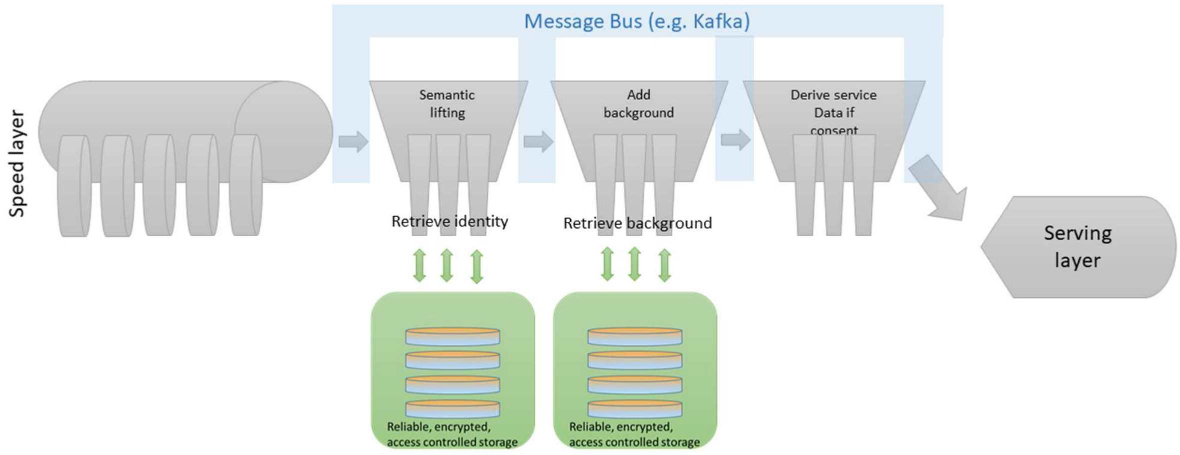

Message Bus (e.g. Kafka)
Speed layer
Semantic lifting
Add background
Derive service Data if consent
Retrieve identity
Retrieve background
Serving layer
Reliable, encrypted, access controlled storage
Reliable, encrypted, access controlled storage

Figure 6: SPECIAL Streaming Data Processing

[12] The term batch is used both in the speed and batch layer. However the size of the batch is different. In the speed layer, the batches are roughly in range of a few KBs, where-as for the batch layer, the size ranges from 10's of MBs to GBs.

Streaming data processing requires efficient communication connecting the different processing steps. This is coloured in blue in the figure above. Such communication system must be designed for supporting scaling on volume and velocity. In SPECIAL, Kafka[13] will be used[14] for this purpose. In addition to be a Big Data technology industry standard, this message bus has built-in, secure communication channels and retention policy support. These properties are valuable for the implementation of a data processing system that complies with the GDPR.

3.4 Implementation considerations

In the previous sections we touched upon a data processing ecosystem that can be used to address the key functional and associated technical requirements for policy aware data processing required in order to realise our use-cases.

In addition to these requirements, all our components in the ecosystem (and the ecosystem as a whole) must adhere to a general requirement of data security. More on our approach to identify the privacy threats and the possible mitigation strategies are found in the next section.

Besides these, the following considerations are to be taken into account when realising CTC components

- *Storage*: The amount of data that needs to be stored can become easily voluminous. Parameters such as the number of data subjects, the number of consent requests and the number of data processing steps, have a multiplicative effect.
- *Scalability*: Because of the multiplicative effect is it important that the SPECIAL architecture can adapt to larger volumes i.e. via both horizontal and vertical scaling.
- *Responsiveness*: The total volume of data should only marginally impact the responsiveness of the services. Creating a single data store will destroy the data locality for some services, impacting the responsiveness.
- *Robustness & long term applicability:* Since CTC management is bound to a legal obligation, solutions should be guaranteed to work for many years. For personal data, the GDPR calls for a long-term durable solution. If changed, the new system should be capable of importing the existing CTC data.

[13] http://kafka.apache.org/

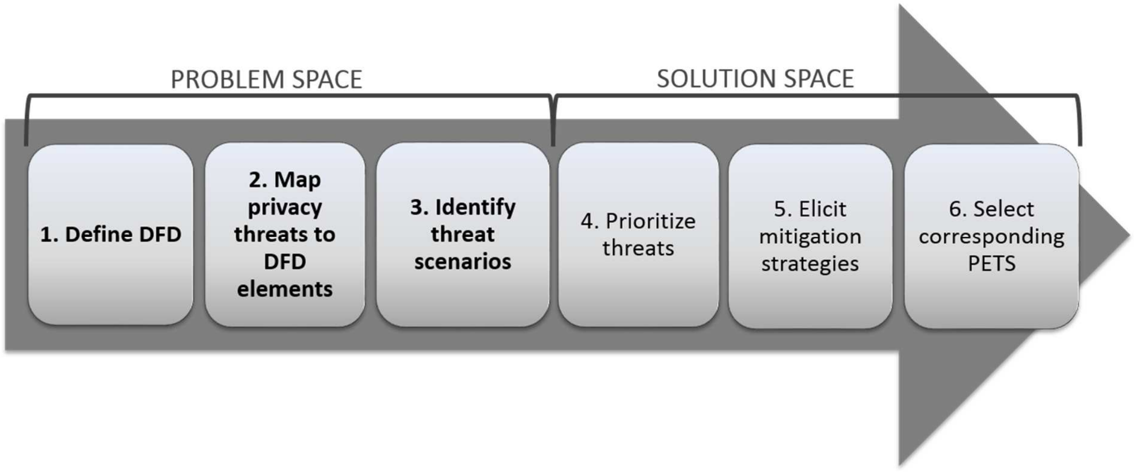
[14] Alternatives are for instance AMQ, Kestrel, JMS, Amazon kinesis.

4 Assessing privacy threats

The SPECIAL project focusses on building consent-awareness and transparency support for data processing systems. The to-be created components themselves are subject to privacy threats. In order to assess these threats and take appropriate mitigation actions, all the software will be evaluated using the LINDDUN[15] methodology. LINDDUN refers to the different treat categories the methodology distinguishes: Linkability, Identifiability, Non-repudiation, Detectability, Disclosure of information, Unawareness, Non-compliance.

This methodology is a threat modelling technique which aids in highlighting the possible privacy treats and the mitigation actions that must be taken. It systematises the development process with regards to privacy treats. Applying LINDDUN results in an overview of the treat status of each component.

The approach, illustrated below, consist of 6 steps of which 3 are applied during the problem definition phase, and 3 steps during the solution design phase.


PROBLEM SPACE
SOLUTION SPACE
1. Define DFD
2. Map privacy threats to DFD elements
3. Identify threat scenarios
4. Prioritize threats
5. Elicit mitigation strategies
6. Select corresponding PETS

In short, the steps are:

1. Define the data flow diagram (DFD) based on the high level description of the system. The modelling entities are *external entities, data stores, data flows,* and *processes.*
2. Map the privacy treats to the DFD. When a privacy treat is acknowledged as short description is given too.
3. Identify the treat scenarios. For each identified privacy treat in the mapping one or more treat exploitation scenarios are designed using a tree representation.
4. Having the scenarios, the next step is to prioritise them using a risk assessment.
5. Next, in order of priority, the treat mitigation approach is defined.
6. Finally, the solutions are detailed by selecting & implementing an appropriate Privacy Enhancing Techniques.

The LINDDUN methodology aids in identifying system wide treats, however some of the treats might be inherent to the chosen technology. In that case, either the technology must be replaced with a better alternative or SPECIAL has to investigate improvements so that the threat's impact

[15] https://distrinet.cs.kuleuven.be/software/linddun/index.php

is reduced. For instance, identity management is such a topic. To relate entities with each other, each data entity needs an identifier. The scope of the identifiers can be either global or local. For Linked Data, the base data representation formalism in SPECIAL, global scoping is normally assumed. Local scoping is possible, but it is usually less supported by the applications. Indeed, global identifiers have the following benefits: they condense data representation, lead to high reuse, and allow easy identification of entities. However the latter benefit is at the same time a data privacy threat as it makes unlocking sensitive data easy.

4.1 Data privacy threats mitigations

The GDPR not only defines functional requirements (constraints and obligations) for a data processing system, it also states that the software components must be designed with data privacy in mind (called the Privacy by Design principle in GDPR). Software solution providers should apply the best practices at the time and are advised to constantly improve their solutions so that the processed data is handled securely.

Hereunder we present a set of characteristics that will impact the design of the to-be developed software components and pilot setups.

4.1.1 Authentication & Authorisation

Personal data should only be accessible after identity of the data requester is confirmed. Authentication is the process which establishes this identity confirmation. Authorisation is the process to confirm whether the identified user has the right to execute a service.

During the past decades access control for end-user facing solutions has reached a maturity. Today two industry standards are widespread:

- OAuth2[16] & OpenId Connect[17]
- SAML2.0[18]

Both offer equivalent functionality through equivalent authentication & authorisation flows.

Authentication & Authorisation is a necessary requirement for the externally accessible interfaces such a user interfaces, but also it is important to consider it for the internal data exchange processes. A multi-tier architecture, integrating an authentication & authorisation layer on the internal APIs creates additional security against unwanted penetration. Such a multi-layer approach is decreasing the likelihood that the impact of a data breach is large, but at the same time it may come at an additional operational cost.

4.1.2 Encryption

A second measure to increase the data security is the application of data encryption. Encryption is the process of encoding the information so that it is only readable by trusted parties having the key to access it.

Encrypting data addresses scenarios such as:

- Unintended disclosure of the data to other system users, in particular users with high rights such as system admins

[16] https://oauth.net/2/

[17] http://openid.net/connect/

[18] https://wiki.oasis-open.org/security/FrontPage

- Easy disclosure of the data in case the system has been hacked or if the system is accidently exposed to the public
- Allows to share data over public channels,
- Reduces the risk of receiving tampered data as tampering requires to break into the encryption

The above scenarios correspond to the following common application areas for encryption techniques:

- The data itself
- The storage medium
- The communication channel

For the latter two, we can mostly rely on the application of existing industry standards and best practices. Encrypting/decrypting on the fly of data being stored in a storage medium is a common offering by cloud providers[19]. Communication channels such as HTTP & telnet, are being replaced with their secure variants HTTPS [20]and ssh[21].

For SPECIAL, encryption of the data itself is more of an open problem. Linked Data is commonly used and exchanged as plain text. The Linked Data ecosystem does not have a built-in approach in which the data represented in RDF is encrypted and stored. Research into the creation of encrypted RDF is therefore part of the research objectives of SPECIAL. Our work on *Self-Enforcing Acccess Control for Encrypted RDF* [22] demonstrates how predicate-based encryption can be applied to realize fine-grained access control on triple patterns over encrypted RDF datasets. In the course of the project, these techniques will be integrated in the SPECIAL platform.

4.1.3 Anonymisation

Anonymisation is a technique turning a source dataset into an equivalent dataset with respect to some properties so that the identifiable real world data subjects present in the source dataset cannot be derived from the anonymised dataset. According to legal interpretation of the GDPR and related legislation, anonymized data can be used more freely. A discussion on this topic can be found in Deliverable 1.2, from page 13 onwards.

However, based on the use case descriptions and the presented SPECIAL ecosystem, the application of anonymisation will be rather limited in the project. Consent management requires access to the identity of the data subject so that data processing steps can apply the consent as requested.

Moreover, none of the state-of-the art anonymization techniques realises full anonymisation[23], but at most a pseudo- anonymisation, the project will not rely on this risk mitigation technique to be GDPR compliant. At most, the pseudo-anonymisation will be used as an additional obfuscation reducing the impact of a privacy data breach.

[19] A description for the Azure cloud storage is found here: https://docs.microsoft.com/en-us/azure/storage/common/storage-service-encryption

[20] https://www.w3.org/2001/tag/doc/web-https

[21] https://www.ssh.com/ssh/protocol/

[22] Self-Enforcing Access Control for Encrypted RDF, Javier Fernández, Sabrina Kirrane, Axel Polleres and Simon Steyskal, Proceedings of the 14th European Semantic Web Conference (ESWC2017), 2017

[23] See Deliverable D1.2, p 16.

4.1.4 Purpose based data storage & data access

The GDPR stresses the aspect that the data is only to be stored, used and shared for the purpose consented to. This legal perspective can inspire the technical perspective on how the data is stored and made accessible.

Ideally a data processing environment should only request data for which it has the permission to get it at the time it needs. Often, still today, application engineers assume that access to the data is granted all the time for the entire duration of a data processing. This simplifies the implementation. Another common activity in software projects is the creation of a developer friendly uniform way to get access to all the possible needed data for the data processing. Usually, the complexity that not all information about a resource is shareable, but only some of its properties when some conditions are met, is ignored. Both attitudes enlarge the risk for unwanted disclosure of data.

Such scenarios are the motivation for research on designing data access control models capable to express access to data based on various properties (relating to the subject, resource or the environment). Attribute Based Access Control (ABAC) and Context Based Access Control (CBAC) are the predominant works in this area. Within these access control models, the purpose for which the consent has been given can be considered as an attribute or part of the context to determine the permission to be accessed.

For SPECIAL, this means investigating if such expressive access control models can be applied to Linked Data. In the Semantic Web journal paper *Access control and the Resource Description Framework: A survey* [24], the authors provide an extensive survey on access control for the Semantic Web. The survey lists a substantial amount of approaches and relevant research, but has to conclude that research investment is necessary to close some important gaps. Nevertheless, enhancing our SPECIAL Linked Data with additional access control features inspired from the research are beneficial and will reduce the risks for unwanted disclosure of personal data.

[24] Sabrina Kirrane, Alessandra Mileo, Stefan Decker:

Access control and the Resource Description Framework: A survey. Semantic Web 8(2): 311-352 (2017), http://www.semantic-web-journal.net/system/files/swj1280.pdf

5 User Interface Preliminary Analysis

The CTC dashboard for the data subject is intended for use in the context of exercising data privacy rights granted by the GDPR. For SPECIAL, these rights are summarised as follows:

- Execute the right of access
- Obtain information about involved processors
- Request rectification or erasure of data
- Consent review and withdrawal

In this section, we present some initial considerations on both the right to access and to obtain information about the involved processors. A more detailed analysis including the provision of concrete recommendations are left to *Deliverable D1.8 Technical requirements V2.*

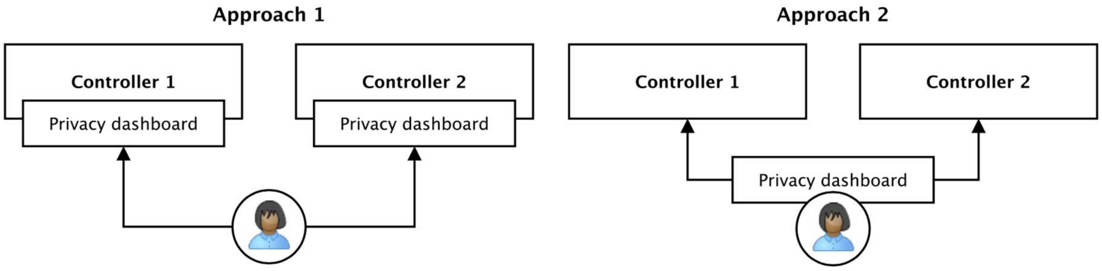


Figure 7: Architectural alternatives for the deployment of the privacy dashboard. Either as single point to manage all controllers, or as data privacy management tool for every controller separately.

We, ideally, envision one privacy dashboard to manage all privacy rights with respect to all controllers a data subject interacts with. That being said, *Figure* **7** shows two different approaches. *Approach 1* requires each controller to deploy and operate their own instance of the interface, which the data subjects can access individually, while *Approach 2* allows data subjects to access one instance of the dashboard that contains information with respect to all controllers that they deal with, however it is more challenging from an architectural perspective.

Execute the right of access

To execute the right to access, all personal data has to be presented to the user **Error! Reference source not found.** ex-post like Siljee's personal data table[25]. This enables answering the question: What data about me did the controller in question collect? The most challenging aspect of this task is to realise the visualisation of huge amounts of diverse data.

One option would be to categorise data based on some taxonomy. One potential categorisation is proposed by Schneier[26], who developed a data privacy taxonomy for social networks. A brief description of the categories is given below:

- **Service data** is any kind of data that is required to provide the service in question (name, address, payment information).
- **Disclosed data** is any data that the data subject intentionally provided on the own profile page or in their posts.

[25] SILJEE, Johanneke. Privacy transparency patterns. In: Proceedings of the 20th

European Conference on Pattern Languages of Programs. ACM, 2015. S. 52.

[26] SCHNEIER, Bruce. A taxonomy of social networking data. IEEE Security & Privacy, 2010, 8. Jg., Nr. 4, S. 88-88.

- **Entrusted data** is any data that the data subject intentionally provided on other users' profile pages or in their posts.
- **Incidental data** is any kind of data provided by other users of the service about the data subject (a photo showing the data subject posted by a friend).
- **Behavioral data** is any kind of data the service provider observes about the data subject while he or she uses the service (browsing behavior).
- **Derived data** is any kind of data derived from any other category or data source (profiles for marketing, location tracks, possible preferences).

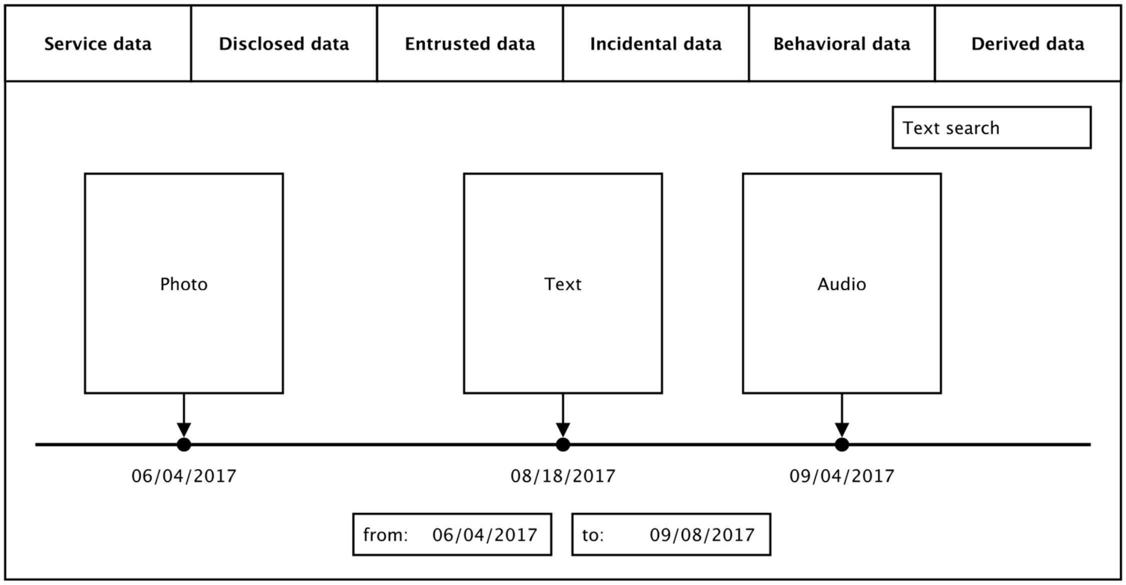


Figure 8: A draft of the interface showing the separation of the different data categories, the text search functionality, the data ordered according to the time of its processing, and the time range selector.

In the wireframe design presented in *Figure **8*** the taxonomy proposed by Schneier is used as a means of categorising personal data.

A complementary approach is to enable users to drill down by limiting the presented data based on some time criteria, for example by year, month, week, day or hour. Such functionality could be implemented, for example via a timeline.

Additionally, a text search function could be offered so that the user can search for specific data items within one of the categories. A search function would enable users to easily find answers to questions like: Did I ever reveal the name of my bank in one of my posts or comments? Could my search engine provider infer the political party I support?[27]

While, filter functions could be offered to the data subject in order to make it easier for the data subject to search for specific types of personal data. For example, a user may want to find out whether a certain controller has audio records of him or her.

Obtain information about involved processors

[27] It is worth noting that Natural Language search is outside of the scope of the SPECIAL project

To use the dashboard to obtain information about involved processors, a graph (see *Figure* **9**) could be displayed that shows the user data flows between controllers and involved processors. In real-life scenarios, many processors are often involved in the processing of personal data. There can be multiple controllers as well (so-called joint controllers[28]). Depending on the number of involved processors in processing the data subject's personal data, the complete graph can be shown as whole or processors can be clustered into groups according to their business domain, for instance.

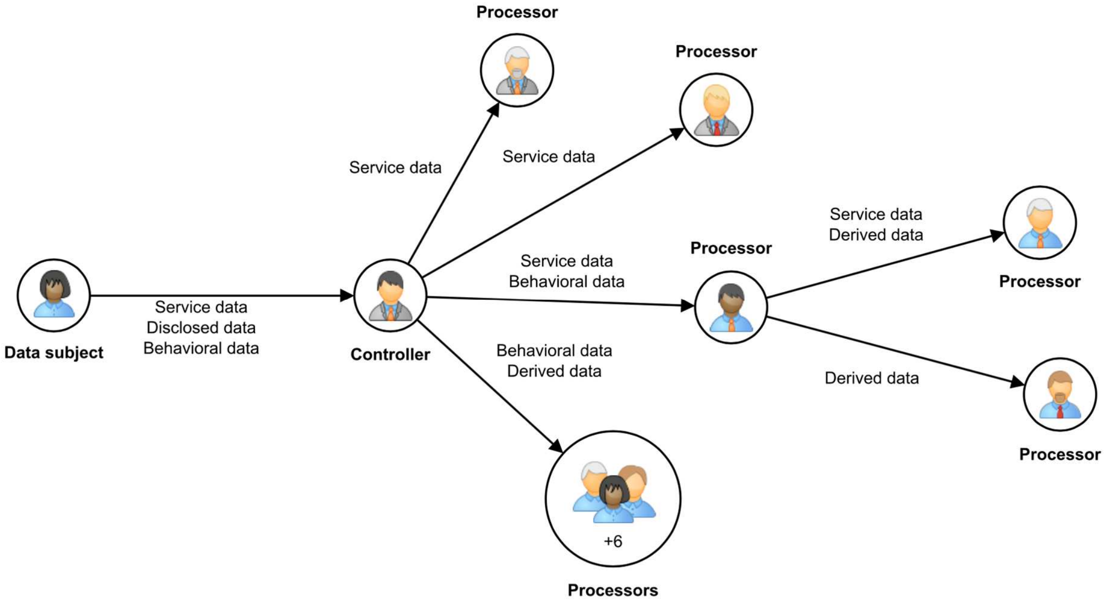


Figure 9: A graph representing external data flows to involved processors. In addition, the forwarded data is categorized to give the data subject information on which data is transmitted.

[28] GDPR art. 26

6 User stories

This section is the next step further eliciting the requirements for components within the SPECIAL ecosystem. We will describe a collection of user stories defining target characteristics of our SPECIAL platform. We have user stories for each of the identified stakeholders, but also for the objectives of the SPECIAL project with specific attention to the impact of the hacking challenge.

During the project execution, the listed user stories will be further elaborated. It is expected that new ones can be added. The presented order does not express a priority. In collaboration with all project partners, a prioritisation will be made, leading to a project roadmap that tackle the user stories.

6.1 User stories for stakeholders

6.1.1 Data Subject

As a data subject I want to …

- browse my personal data
- explore the policy definitions
- adapt my personal data
- adapt my consent
- request transparency data
- browse the transparency data
- request to be forgotten (erase my data)
- get notifications on status changes for my requests
- have my data securely stored
- have my consent honoured
- export the consent I have given
- import consent from a third party
- have secure access to the portal

6.1.2 Policy administrator

As a policy administrator I want to …

- browse the policy definitions
- define policy definitions
- explore the usage of policy definitions
- have a change to be propagated to all
- have secure access to the portal

6.1.3 Auditor

As an auditor I want to …

- browse the policy definitions
- explore the usage of consent

- see the consent of a data subject
- explore the transparency data w.r.t. the given consents
- explore the catalogue of data sources used for the service, with their legal information (licenses, consents, ownership, etc.)
- have secure access to the portal

6.1.4 Company (Service Provider)

As a service provider, I want to …

- be able to create a business valuable service without violating the legation and the data subject's consent
- share & sell the resulting data for which consent was provided
- have a secure data processing solution with minimal personal data discloser risks
- have a reliable way to implement the right to be forgotten

As a service developer, I want

- have easy, secure & standard access to the consent of a data subject
- have easy, secure & standard way to log the data processing provenance trail
- have the hooks to implement the right to be forgotten

6.2 Service consumer

As a service consumer I want to …

- have the guarantee that the service is based on trustworthy, reliable acquired data

6.3 User stories for SPECIAL objectives

We add some key user stories covering the perspective of the SPECIAL project. In contrast to the stakeholder user stories, these reflect the research and technical ambitions the project has.

As SPECIAL consortium, we want to …

- have a simple deployable & development environment for the platform and its pilots
- have a privacy treat analysis for the components of the SPECIAL platform
- have a domain independent consent ontology
- have a domain independent policy ontology
- have a domain independent transparency ledger
- have a reliable, trustworthy policy engine protected against privacy threats
- have a reliable, trustworthy transparency ledger engine protected against privacy threats
- have 3 pilot instances of the SPECIAL platform, each of them corresponding to a use case

6.4 Hacking challenge

One of the objectives of the project is to setup a public hacking challenge to evaluate the SPECIAL platform. Instead of merely creating a public instance and hoping the anonymous internet society finds it and attempts to penetrate it, it is our intent to create a number of hacking challenges around the privacy protection measures the platform has.

An example of such a challenge is trying to break into the policy engine with the intent to alter the response on the query if there is consent. Obviously, if that is possible, the policy engine's responses cannot be trusted and hence data processing relying on the consent is untrustworthy.

At this moment, these scenarios are not fixed. This is future work that will be collected during the next phase in the project. These scenarios will become key user stories.

Setting up a hacking challenge imposes an important milestone in the project with respect to the technical readiness of the involved components and data. At the launch of the hacking challenge, the components are to:

- install and deploy easily on the hackers' local infrastructure (since it is not our objective to have the hacker challenge our project's cloud infrastructure),
- have the desired functionality,
- have a documented list of unimplemented features/weaknesses (to avoid reporting issues which we are aware of)

and must there be representative syntactic data available. It might be required to have a syntactic data generator in order to support the execution of the hacking challenge.

Aside from the technical requirements the success of a hacking challenge depends on a good communication strategy and expectation management. The communication strategy must initiate enthusiasm in the targeted community. This may be achieved with additional motivation by for example offering a prize.

7 Software & system design principles

In deliverable D3.1 the initial deployment of the SPECIAL platform has been described. It mainly focussed on the deployment and implementation aspects of the SPECIAL platform reusing the BDE platform, its extensions and experiences.

The a-priori choice for the BDE platform is motivated by the following arguments:

- its functional capability to create easily complex big data ecosystems combining a wide variety of technologies and operational contexts,
- its integration with Linked Data processing, and
- the experience for SPECIAL project partners with it.

Where-as the latter capitalizes on the reuse potential by the project partners, the first two are prerequisites for the SPECIAL ecosystem. During the project SPECIAL will extend / adapt the BDE platform to its needs, creating a privacy aware BDE platform.

In the next section, we list a number of technical design principles that will be applied for the development of the SPECIAL platform.

7.1 Operational environment

Principle 1) Automated system rollout as much as possible.

Using system deployment descriptions such as Terraform (system resources layer) and Docker Compose (services layer) the roll-out of an application becomes reproducible and reliable. Because the description is stored in a source control repository, changes over time and variants can be maintained without the need of having them actively running. The consumption of system resources can then be dedicated to the active developments.

Principle 2) Cloud enabled by design

Our platform has to be hardware and Operating System neutral as much as possible. Using a service abstraction layer (i.e. Docker) addresses one part. Additionally the setup has to be decoupled from the local file system. Only then will the system be completely cloud enabled and runnable independently.

7.2 System architecture

Principle 3) Modular design, by preference following the micro-services pattern

A micro-service design is the idea to create a system from the integration of a collection of services, each with a dedicated purpose. This approach enables a scaling potential for the system: if one service is in high demand, adding new services of the same kind is a straightforward action. In addition, it allows to focus the development effort. The approach has proven results in the design for end-user facing software. A similar approach can be found for the data processing: the design techniques for the speed layer of the lambda-architecture (stream processing) recommend to create small dedicated data processing steps that are combined into one larger data processing pipeline.

Principle 4) Reuse best practice standards for well-known technical challenges

As already mentioned in section 4, many privacy threats have industry supported mitigation strategies. Therefore, unless they are not sufficiently appropriate or adequate, it is our strategy to apply the best practices as much as possible.

7.3 Component interaction

Principle 5) Payload data is preferable in the form of RDF, json-LD or json.

Although the use-cases indicate that data from various sources with a multitude of formats are processed to create the desired value-adding services, it is our intent to keep the heterogeneity as much as possible under control by using preferable RDF, json-LD or json as payload data representation. Where-as RDF and json-LD are highly compatible with each other, json requires additional semantical lifting. This lifting can be defined by adding an LD context to the json payload. Thus, although not technically imposed that the data is exchangeable, these 3 data representation formats can form a uniform data landscape.

When a component does not comply with this preference, it may be required to create a dedicated payload translation layer for the component. To some extent, semantic lifting acts as such wrapping.

Principle 6) The data-exchange channels are secure.

The payload data has to be exchanged between the services. Most importantly is that the used data exchange channel is secured against penetration: HTTPS, secured database connectors (ODBC, JDBC), secure file access and a secure message bus (Kafka) are the preferred choices.

8 Conclusions

We have provided a global technical overview of the SPECIAL ecosystem, identifying the key stakeholders with their main user stories, and the high level software design principles. With this deliverable and the other initial requirements analysis deliverables, a next phase of the project starts.

The common understanding allows the project to define an implementation and research roadmap for the following months. A project roadmap consisting of several iterations will be created. The first iteration will create the initial SPECIAL platform; the subsequent ones will extend and enrich the SPECIAL platform with new or improved functionality and insights.

Throughout this collaboration, the requirements will be made more detailed and collected in the form of Deliverable *D1.7 Policy, transparency and compliance guidelinesV2*, which will be delivered at the end of month seventeen.