A framework for GDPR compliance in Big Data systems *

Mouna Rhahla[1,2], Sahar Allegue[1,2], and Takoua Abdellatif[1]

[1] University of Carthage, Polytechnic School of Tunisia, SERCOM, Tunisia
Takoua.Abdellatif@ept.rnu.tn
[2] Proxym-Lab, Proxym-IT, Sousse Tunisia
Forname.name@proxym-it.com
http://www.proxym-group.com

Abstract. The verification and implementation of the GDPR regulation that aims at protecting European citizens' privacy, is still a real challenge. In particular, in Big Data systems where data is of huge volume and heterogeneous, it is hard to track data evolution through its complex life cycle ranging from collection, ingestion, storage and analytics. In this context, from 2016 to 2019 research has been conducted and security tools designed. However, they are either specific to special applications or address only partially the regulation articles. In order to identify the covered parts, the missed ones and the necessary metrics for comparing different works, we propose a framework for GDPR compliance that identifies the main components for the regulation implementation. Based on this framework, we compare the main GDPR solutions in Big Data domain and we propose a guideline for GDPR verification and implementation in Big Data systems.

Keywords: The General Data Protection Regulation · Big data analytics · Security · Privacy.

1 Introduction

GDPR [1] sets new standards on security through 99 articles and 173 recitals and aims to protect the rights and freedoms of natural persons. Every organization that deals with data has to comply with GDPR, to protect these rights and to be accountable while improving business models [2]. Accountability aims at demonstrating how controllers comply with data protection principles. Each organization must answer the following questions: what information is processed? why? how and where is data stored? who can access it and why? is it up-to-date and accurate? how long will you keep it for? how will it be safeguarded and how do you reach accountability?

Answering all these questions guides system designers and developers to verify and implement GDPR compliance. It involves several tasks, namely the reading of the regulation articles, the knowledge extraction and the characterization

* This project is carried out under the MOBIDOC scheme, funded by the EU through the EMORI program and managed by the ANPR

of the regulation concepts following the system specification. Mapping the regulation articles and concepts to software can be a tedious and confusing task. Consequently, a model is needed to describe the concepts captured from different GDPR documentations, to extract GDPR principles, actors and dependencies. The model should identify the main building blocks to be implemented for GDPR compliance verification and implementation. To reach this goal, our work promotes an improved understanding of the legal, organizational and technical concepts present in GDPR documentation and classifying, the different related works in a clear manner.

In the last few years, topics about GDPR have been widely discussed across a range of academic publications and industry papers from different theoretical and practical perspectives, including numerous implementations and design concepts for GDPR compliance [35]. These works are still in their infancy and mature and formal tools that implement GDPR articles are still missing especially in the area of big data analytics. The term big data analytics alludes to the entire data management life cycle from ingestion and storage to analysis of high volumes of data with heterogeneous format from different sources. As presented in figure 1, the reference architecture of big data systems covers 4 main layers: data sources, ingestion, processing, storage and services. At service and processing layers, sophisticated algorithms are being developed to analyze such large amount of data to gain valuable insights for accurate decision-making, detecting unprecedented opportunities such as finding meaningful patterns, presuming situations, predicting and inferring behaviors. Due to the large data volume and the complexity of processing, tracking data dependencies and privacy verification are challenging. For this purpose, data security and governance layer is a cross-layer generally used for data security and management. It represents a key part of the system in implementing GDPR principles. However, accurate metrics are needed to evaluate the GDPR coverage of this management layer.

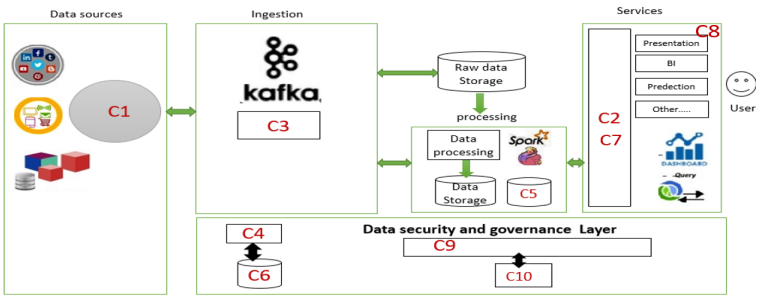

Fig. 1. A classical Big Data system following a reference architecture

Recent academic and industrial tools [14] implement privacy by design translating automatically the privacy policies to software and providing accountability. However, these works address only partially GDPR principals such as purpose limitation, data minimisation, storage limitation, transparency or security [11][12]. Other works concentrate on particular articles of the regulation and do not cover all of them [10] like the right to data portability, the right to be forgotten, the access right or the right to be informed [21][22]. Also, these works generally address one particular type of data source (logs, IoT sensors or classical SQL data bases). It is not clear how to apply proposed solutions that consider uniform data, to big data architectures with multi-channel data sources, different purposes and intensive processing. Consequently, we still lack guidelines to verify GDPR compliance and to implement the regulation in a Big Data context. As a starting point, in order to address this issue, a comprehensive overview of the regulation and a common understanding of its key concepts are necessary. Later, the study of recent works on privacy and GDPR allows the identification of the main building blocks for GDPR compliance verification. Thanks to both theoretical and the implementation analysis of the regulation, we propose a framework with well-defined components to implement the regulation. According to these components, we situate the different works carried out on GDPR in the domain of Big Data. Furthermore, we provide an overview of how to use the framework to assist Big Data engineers and system administrators to build GDPR-compliant systems and applications. As a use case, we consider the example of Big Data client segmentation application and we propose an implementation of the framework components to have the application GDPR-compliant. This paper's contributions can be summarized into the following points:

- A framework for GDPR compliance in Big Data systems.
- A classification of the state of the art conducted on GDPR solutions implemented between 2016 and 2019 in both academic and industrial areas.
- A guideline for GDPR-compliance in Big Data applications based on the framework components.

This paper is structured as follows. Section 2 is an overview of GDPR principles and its main actors. In section 3, we describe a framework for GDPR in Big data systems. In section 4, we present the framework used to classify GDPR-related works in Big Data. Section 5 describes the framework usage to implement GDPR compliance in a Big Data application. Finally, section 6 gives a summary of the main findings of this paper and highlights new opportunities for future work.

2 Analysis of GDPR principles and actors

The GDPR regulation aims at delivering a harmonized, consistent and high-level data protection across Europe. It has 99 articles grouped into 11 chapters. In those chapters, it addresses a set of principles, actors and obligations.

2.1 GDPR principles

GDPR regulation sets out seven key principles for the processing of personal data stipulated in Article 5 [13]. They can be summarized into the following points:

- **Lawfulness, fairness and transparency:** "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject." Art.5(1)(a). More detailed provisions on lawfulness are set out in Articles 6 to 10 and detailed transparency obligations set out in Articles 13 and 14. Lawfulness, fairness and transparency may overlap, but all three must be satisfied in a system. In fact, it is not enough to show your processing is lawful if it is basically unfair to or hidden from the data subject concerned.
- **Purpose limitation:** "Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes." Art.5(1)(b). This principle aims to make sure that you are clear and open about your reasons for collecting personal data and that what you do with the data is in line with the reasonable expectations of the data subject concerned. Specifying your purposes from the beginning helps you to be responsible and accountable for your processing.
- **Data minimisation:** "Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed." Art.5(1)(c). The minimum amount of personal data is revealed to satisfy the application purpose. The accountability principle means that you need to be able to prove that you have appropriate processes to make sure that you only collect and hold the personal data you need.
- **Accuracy:** "Personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay."(Art 5(1)(d)). Here we find clear links to the right to rectification, which gives data subjects the right to have incorrect personal data corrected.
- **Storage limitation:** "kept 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; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organizational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject." Art.5(1)(e). So, even if you collect and use personal data fairly and lawfully, you cannot keep it for longer than you actually need it. GDPR does not set specific time limits for different types of data. This is up to you, and it will depend on how long you need the data for your specified purposes. Once information is no longer needed, personal data must be securely deleted.

- **Integrity and confidentiality:** "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorized or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organizational measures." Art.5(1)(f).
- **Accountability:** "The controller shall be responsible for, and be able to demonstrate compliance with Art 5(1)." Art.5(2). The accountability principle requires you to take responsibility for what you do with personal data and how you comply with the other principles. You must have appropriate measures and records in place to be able to prove your compliance.

2.2 GDPR actors and their dependencies

There are five main actors in the regulation [13]:

- **Data Subject:** an identified or identifiable natural person, directly or indirectly, by means reasonably likely to be used by the controller or by any other natural or legal person.
- **Controller:** a natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes, conditions and means of the processing of personal data. It ensures compliance with GDPR principles related to the processing of the personal data (Accountability), implements data protection policies and data security measures, carries out data protection impact assessment (DPIA) for high risk processing, informs data subjects on their rights, in case of personal data breach, notifies the supervisory authority within 72 hours and transfers personal data to third country or international organization, per specific safeguarding provisions.
- **Processor:** (a person or a legal entity) processes personal data on behalf of the Data Controller, specifically: collects personal data online through registration, contact forms, email or digital payments and invoicing, stores, uses, records, organizes, retrieves, discloses, deletes the collected personal data on behalf of and under the instructions of the Data Controller and creates inventories for all above mentioned data processing categories.
- **Data Protection Officer (DPO):** (a person or a legal entity) manages and supervises all data protection activities, specifically: monitors compliance to GDPR's personal data protection and security provisions and cooperates with the supervisory authority.
- **Supervisory Authority (SA):** Article 46 states that supervisory authorities are responsible for monitoring the application of this Regulation and for contributing to its consistent application. The independent public authority is responsible for monitoring regulated entity compliance with GDPR.

Figure 2 shows the main actors and their relations following the regulation articles. For example in figure 2, the data subject can declare his consent to the controller (Art.4). He can also request data from the controller (Art.12). On the other hand, the controller provides information to data subjects (Art.12) and communicates data breaches to them Art.34).

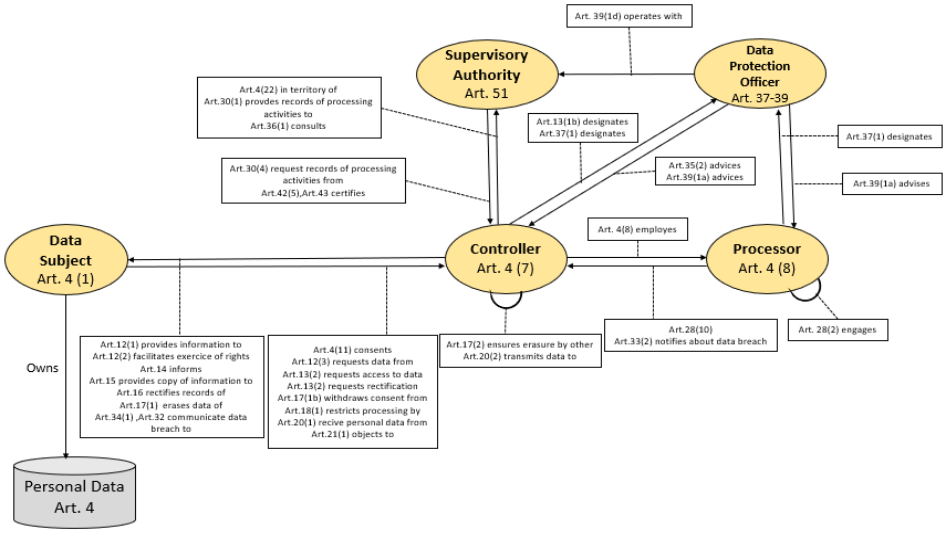

Fig. 2. GDPR main actors and their relations following GDPR articles

3 A GDPR framework for Big Data systems

In order to satisfy GDPR principles, data has to be tracked during all its life cycle and the processing stages. Furthermore, a governance layer is required for the controller.

For a GDPR framework definition, two complementary approaches are followed:

- A top-down approach. We start from GDPR documentation to define a model serving as a language for describing GDPR obligations and requirements in terms of principles, actors and actions. This is what we have presented in previous sections.
- A bottom-up approach. It is used for complementing the previous model by providing a more fine-grained set of components, based on the analysis of relevant related works that will be presented in the next sections.

Figure 3 represents the overall architecture of our proposed framework. For each GDPR actor, we define its main components.

We have set up 10 components (from C1 to C10). Each Ci respects a set of properties as follows:

- **C1:** Data Subject (DS) that represents the data sources. A DS puts restrictions on the use of its collected and exploited data. In the context of Big Data, three Vs are defined as component properties: Volume, Velocity and Variety.

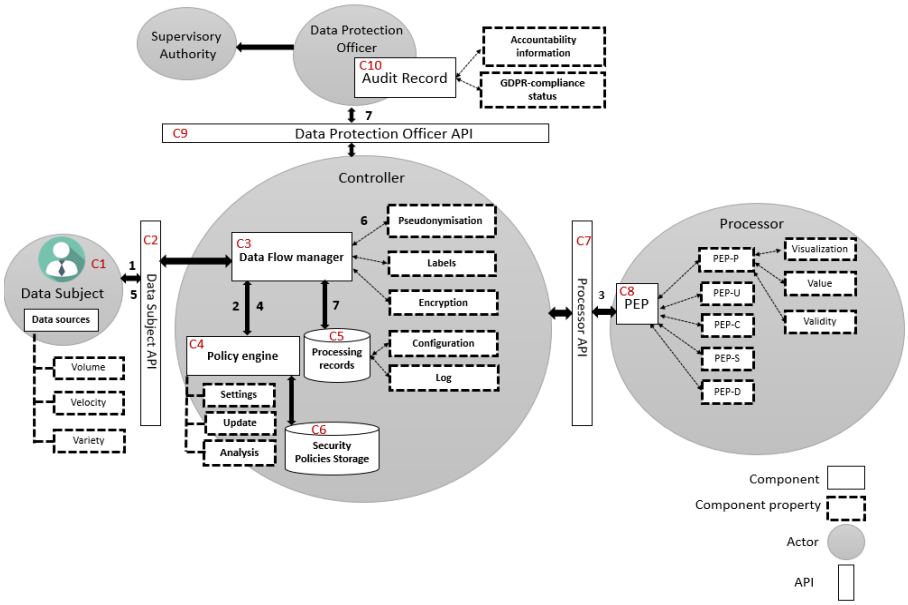

Fig. 3. The GDPR framework architecture

- **C2:** DS API. It provides a tracking dashboard and notifications management interface to the DS. A view of the security control and sensitive data is provided. Furthermore, this API translates the consent and restrictions of the DS from human language to machine language. These restrictions can provide purpose limitation, storage limitation, data minimisation and access control.
- **C3:** Data Flow manager. It allows the DS to define or modify his security policies and his data. It is also used to receive and send data and/or notifications to different components. The Data Flow manager component provides a set of properties: pseudonymisation for data minimisation, labels for data tagging and tracking, and encryption for integrity and confidentiality. Two types of data can be returned to the processor: either pseudonymous data or encrypted data for the processing which have to keep the DS identity secret.
- **C4:** Policy engine. It receives queries from users and as a result, a decision is returned to give the user access to collect data or not. It provides three properties: policy settings, update and analysis.
- **C5:** Processing records. It is a component where all operations applied to data are stored to satisfy transparency principal. Two types of records are stored: logs and configurations.
- **C6:** Security Policies Storage. This component interacts with C4. Indeed, policies defined or updated by C4 are stored in C6.

- **C7:** Processor API. It allows the controller to interact with the processor and more specifically the PEPs (Policy Enforcement Points).
- **C8:** Policy Enforcement Point. It is the point where the processor executes a specific purpose. PEP tasks or properties are: Processing (P), Collection (C), Usage (U), Storage (S) or Distribution (D). Every PEP must take into account Big Data Vs constraints such as: Visualization, Value and Validity for PEP-P (PEP-Processing).
- **C9:** Data Protection Officer API. It helps the controller interact with the Data Protection Officer to demonstrate compliance and accountability.
- **C10:** Audit Record. It records information and data to provide accountability and to track GDPR-compliance status.

In the following table, we present the matching of GDPR principles against the framework components. It shows that the framework covers all principles. Also, this table can guide systems designers and developers in identifying the components to implement for a given purpose.

GDPR principles	The framework components
Lawfulness, fairness and transparency	C2, C3, C5 and C9
Purpose limitation	C3, C4 and C6
Data minimisation	C3, C4 and C6.
Accuracy	C4 and C6
Storage limitation	C4 and C6
Integrity and confidentiality	C3, C4 and C6
Accountability	C10

4 GDPR-oriented solutions in Big Data systems

As the GDPR regulation was initiated in 2016 and came into effect on the 25th of May 2018 as a new privacy regulation for the European Union (EU), related work on GDPR compliance is limited. The proposed works have been investigated from either the perspective of jurisdiction or information system implementation. In our study we collected 6 major refereed academic databases (Education Resources Information Center (ERIC), JSTOR, Google Scholar, ACM publication, IEEEXplorer, Wiley Online Library and SpringerLink) in the field of Computer science indexed by Journal Citation Reports using the search terms "EU GDPR" and "EU General Data Protection Regulation". This resulted in 170 papers to be included in our study. After an analysis of the papers title and abstract, the number was reduced to 62 papers.
We applied a set of selection criteria as follows:

1. Papers on GDPR for organization or focusing on personal data protection.
2. Papers providing experimental or empirical studies from actual observations and case studies with private data.
3. Papers presenting a new design for GDPR compliance.

The selections resulted in 62 peer-reviewed publications that fit the criteria above. A high number is reached in 2018 and the number of works still increasing in 2019.

These related works are divided into 4 main categories: (1) Jurisdictional analysis of GDPR, (2) Academic solutions, (3) Industrial tools and (4) Apache Big data technologies and tools. The next sections summarize these four categories.

4.1 Jurisdictional analysis of GDPR

Several authors and organizations have analyzed GDPR and privacy by design [35]. However, they only provided documentation and support for the law without providing any framework or guidelines to apply this law in company's projects or products [13] [25] [26][27]. Other research papers focus on specific sectors such as healthcare [3] [4] or specific purposes like storage compliance with GDPR [5]. Although interesting for better understanding of GDPR, they do not come up with tools or implementations of GDPR articles. They rather analyze and discuss the impact of GDPR on the considered specific fields and sectors [6][9][41] or discuss the feasibility of the privacy by design principle [42].

4.2 Academic GDPR solutions

In the past three years, many authors have worked to provide privacy tools for GDPR. These tools partially cover partially GDPR principles. In [21] a privacy tool is developed for DNS applications. Indeed, some DNS data may reveal personal data (resolver IP addresses and domain names looked up by users). A thorough and transparent approach to privacy protection in this field is considered vital to the Dutch economy and society. This work targets DNS data and only the GDPR transparency principle is treated.
Authors in [10] propose privacyTracker, a GDPR-compliant tool that covers data tractability and transparency. They implement some GDPR rights such as data portability and right to erasure. For GDPR accountability in IoT systems, an IoT Databox model is proposed providing the mechanisms to build trust relations IoT [22]. In a previous work [23], we proposed a GDPR Controller for IoT systems where security, transparency and purpose limitation are implemented. In A4Cloud project [14], a tool-set is composed of eleven tools implementing transparency, privacy, trust and technology for tracking and controlling data [12][31] [32] as well as for policy management [33].
In [24] authors present TagUBig - Taming Your Big Data, a tool to control and improve transparency, privacy, availability and usability when users interact with applications. For IoT system, ADvoCATE [11] allows data subjects to easily control consents regarding access to their personal data. The proposed solution is based on Blockchain technology. Juan Camilo proposed another Blockchain-based solution to implement consent in GDPR [8]. This work provides the data subjects a tool to assert their rights and get control over their consents and

personal data. In [43], authors discussed how static program analysis can be applied to detect privacy violation in programs. The solution is based on classical information flow control techniques, tainting and backward slicing. Although important, the solution addresses a limited part of data control requirements in GDPR.

4.3 Industrial GDPR tools

In addition to academic papers, many industrial security tools have been proposed. We identify some of them in this section:

- **The Absolute Platform:** This tool provides visibility and control. It addresses GDPR prerequisites by observing and verifying PII (Personally identifiable information), avoiding data breaches and automating remediation [15].
- **Alien Vault USM:** This tool helps to detect data breaches and monitor data security. The unified platform centralizes essential capabilities like asset discovery, vulnerability scanning, intrusion detection, behavioral monitoring, log management and threat intelligence updates [16].
- **BigId:** This tool assures data minimisation through duplication discovery and correlation. It satisfies customer data portability, supports and enables right-to-be-forgotten. In addition, it reveals enforcement of customer consent for personal data collection, data residency flows and risk profiling with breach notification windows [17].
- **BWise GDPR Compliance solution:** this tool helps to build data views, data control and compliance [18].
- **Consentua:** It is a consent choice and control tool that enables users to choose and control over their personal data. It empowers an increasingly trusted and straightforward relationship between the client and the service provider[19].
- **PrivacyPerfect:** This work is composed of a set of tools such as assessment, processing and dashboard tools specially designed for chief privacy officers, reports, legal processing grounds and graphical overviews [20].

4.4 Apache solutions

Apache has developed a set of tools to provide security to Big data systems architectures. These technologies can be used to address parts of GDPR requirements. Here are some popular solutions:

- **Apache Eagle:** Apache Eagle is an open-source solution for identifying security and performance issues instantly on big data platforms like Apache Hadoop and Apache Spark. It analyzes data activities and daemon logs. It provides state of the art alert engine to identify a security breach, performance issues and shows insights [28].

- **Apache Atlas:** Apache Atlas is an open-source solution used for data tagging. It provides open metadata management and governance capacities for organizations to make a catalog of their data assets, classify and govern these assets and provide collaboration capabilities around these data assets for data scientists, analysts and the data governance team [29].
- **Apache Ranger:** Apache Ranger is an open solution that helps developers to enable, monitor and manage the entire data security across the Hadoop platform. The vision with Ranger is to provide a framework for central administration of security policies and user access monitoring [30].
- **Apache Knox:** Knox Gateway provides a single access point for all REST and HTTP connections with Apache Hadoop clusters. Knox API Gateway is structured as a turn around intermediary with thought for pluggability in the areas of policy enforcement, through providers and the back-end services for which it proxies requests [34].

4.5 A comparative study

The comparison of the studied works and tools is tedious work since they come from different communities and target different objectives and variant application contexts. Nevertheless, it is important to situate these works to measure their compliance with GDPR and also for reuse purpose. Indeed, instead of reinventing the wheel, some ideas and implementations can be reused for designed big data systems, even though the work context can be different. Consequently, a global picture of these solutions and their implemented parts of the regulation can be very helpful to big data system designers. Using the defined framework, we can set up the comparative table below. We adopt the following notations:

- ✓: indicates that the component is implemented by the referenced work.
- ∼: indicates that the component is partially addressed by the referenced work and some of its properties are not implemented.
- ×: indicates that the component is not implemented by the referenced work.

In Table 1, the framework components defined in figure 3 are partially implemented in each solution. There are no works that implement all components. More precisely, we can clearly see that the components that are most covered are C2, C4 and C6. They mainly focus on security policy definition and management. This can be explained by the fact that involving people in the definition of their security constraints and the tracking of their data is the focus of many research works even before GDPR was voted. Indeed, providing practical and intuitive API for users, not necessarily experts in security, was considered, for many years, a priority in privacy-sensitive systems like e-health and other IoT systems. On the other hand, C3 addressing data tagging is less implemented compared to other components. Indeed, with data heterogeneity and the multi-sources of Big data, data tagging becomes necessary for tracking and controlling data flows. This technique is less required in small systems with uniform data format and source. Also, C9 and C10 are two components not really dealt with

except in recent few efforts. A new need comes in with GDPR consists in interfacing with the data protection officer and the supervisory authority. More generally, a GDPR audit tool is required bridging the gap between regulation texts and implementations. The challenge consists in creating robust tools that analyse software systems and check automatically that they respect the regulation text. In the opposite direction, translating regulation constraints expressed in human languages to a code respecting regulation constraints is also an interesting axis of research that needs to be developed.

Related works by Category		Our framework Components									
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Jurisdictional analysis of GDPR	GDPR for healthcare [3]	×	×	×	×	×	×	×	×	×	×
	GDPR in Health Clinics[4]	×	×	×	×	×	×	×	×	×	×
	Storage system for GDPR [5]	×	×	✓	∼	✓	∼	×	×	×	×
	GDPR investigation [6]	×	×	∼	∼	×	∼	×	×	×	×
Academic GDPR solutions	Tool for DNS big data [21]	∼	×	×	∼	∼	×	×	✓	×	×
	PrivacyTracker [10]	∼	✓	×	∼	✓	✓	∼	×	×	×
	IoT Databox [22]	∼	✓	∼	×	×	∼	✓	∼	×	×
	GDPR Controller [23]	∼	✓	∼	✓	×	✓	×	∼	×	×
	TagUBig [24]	∼	∼	×	✓	∼	✓	×	∼	×	×
	ADvoCATE [11]	∼	∼	∼	✓	×	✓	×	∼	×	×
	Consent management [8]	∼	×	∼	✓	✓	✓	×	∼	×	∼
	The A4Cloud project [14]	∼	✓	∼	✓	∼	✓	∼	∼	×	×
	Trust and Tracking [12][31] [32]	∼	✓	∼	✓	∼	✓	×	×	×	×
	Policy management [33]	∼	✓	×	✓	×	✓	×	∼	×	×
Industrial GDPR tools	The Absolute Platform [15]	×	✓	∼	✓	✓	∼	×	×	×	∼
	Alien Vault USM [16]	×	✓	∼	×	∼	×	×	×	×	×
	BigId [17]	×	✓	∼	×	×	×	×	×	×	×
	BWise GDPR solution [18]	×	✓	∼	✓	✓	✓	×	×	∼	∼
	Consentua [19]	×	✓	∼	∼	×	×	×	∼	×	×
	PrivacyPerfect [20]	×	✓	∼	×	∼	×	×	×	×	∼
Apache solutions	Apache Eagle [28]	×	✓	∼	∼	∼	∼	✓	✓	×	×
	Apache Atlas [29]	×	∼	∼	×	∼	×	×	×	×	×
	Apache Ranger [30]	×	∼	∼	✓	×	✓	×	×	×	∼
	Apache Knox [34]	×	×	∼	∼	×	∼	×	×	×	∼

Table 1. Comparative Table

5 Using the framework for GDPR-compliance implementation in a Big Data application

In this section, we consider a classical Big Data application: a customer segmentation[36]. Segmentation enables marketers to better customize and maximize sales to diverse audience subcategories. Customer segmentation may be based on

demographic, geographic, behavioral or psycho-graphic criteria and, therefore, generally processes private customer data. As for GDPR actors, the customer represents the data subject, services using data for collection, storage, visualization and analysis represent processors and the controller is represented by the governance layer of the Big Data ecosystem. Classically two kinds of processes are used, RFM and clustering [36]. We can consider them as black-boxes with user data as input and customer classification information as output in the processing layer of figure 1. Customers are grouped based on common characteristics that can be private.

The first step allows the DS to define his restrictions about data access. In our context, this means that he precises via a user interface the persons allowed to access his data, the period of time data will be stored (Storage limitation), the purposes his data will be processed for (purpose limitation) and other access conditions. Component C1 collects and aggregates different types of data about customer transactions. It is deployed at the data source layer and it is connected to C2 so that collected data is aggregated and displayed through the user interface. Indeed, a dashboard is provided by C2 to the customer in order to track his data, to check his timeline and processing records and to check if restrictions adjusted in his configuration settings are respected. For this objective, the solution of [12] can be adopted. Also, C2 can translate defined policies from a human language to security policies as machine language. For that, a tool such as [33] can be integrated.

As a second step, security policies set by the DS are stored into C6 component by component C4. The DS can define, update or delete his consent and/or his data. For that, Apache Ranger [30] is a good candidate for C4 and C6 because it allows different security policies to set up automatically. Furthermore, thanks to its open architecture, Apache Ranger can be extended for specific policy formats and contexts and for third party implementation of policies more specifically, Ranger adopts Attribute Based Access Control (ABAC) [37]. We can add other types of attributes in the PIP (Policy Information Point) so that we can manage the semantics of the policy and even combine different policies to provide new rules. Many works are proposed on ABAC architecture [23] [37] that can be integrated into the Apache Ranger. But work is still needed to evaluate this cryptography overhead on processing time. At regular periods, the application sends a segmentation query that crosses components C3 to C7. The query is evaluated at C4 to check if the application processors are allowed to use the DS data for the segmentation purpose. More precisely, the security policy is evaluated: purposes limitation, storage limitation with the fixed period by the DS, data minimisation and access control restrictions. Then, according to policies, data is filtered at C3 that takes place in the ingestion layer of figure 1. Kafka [39] is classically used as a pipeline for Big Data. Furthermore, Kafka can be a good hub between the different framework components. Thanks to the publish/subscribe architecture, components are notified about new data, new security policies or for other data management purposes. For example, we can choose a topic for notification management, another topic for policy manage-

ment and a topic per data source [23]. Before reaching the processing step, C3 provides a set of possible features to implement the GDPR "Data minimisation" principle. Data may go through labeling, encrypting or pseudonymisation process depending on the DS restriction. The goal is to communicate data only relevant, adequate and limited to what is necessary. Tokenization is an interesting technique to distinguish private data from public [38]. It can be enforced by encryption property using the crypto-engine component [23] so that only processors involved in the segmentation application can have access to data. C5 archives the different processing logs in NoSQL or SQL databases [40] for visualization and for accountability. Indeed, accountability information is displayed to the data Protection Officer thanks to C10.

Although some implementations of GDPR principles are available, integrating them together in a single solution is still an open issue. Interoperability between security solutions and performance overhead have to be measured and evaluated especially in a Big Data context.

6 Conclusion and future work

This paper presents a framework that details the main components for GDPR compliance verification and implementation. The framework allows classifying and comparing different works related to GDPR, particularly for Big Data systems. Furthermore, this framework can be used as a guideline for implementing solutions and products respecting GDPR. Few works address new requirements related to the heterogeneity and multi-source data of Big Data systems like tagging techniques. Also, a lot of work still be needed to address specific interfaces introduced by the regulation like audit tools for the data protection officer and the supervisory authority. The example of a segmentation application illustrates how to use the framework to assist developers and system designers to implement privacy by design and by default and highlight missing work in the state of the art. As future work, we plan to implement the framework components in a real segmentation big data application. The goal is to demonstrate its GDPR-compliance and to evaluate the overhead of GDPR-compliance implementation mainly on the system performances. As a second step, we are very interested in developing audit tools for GDPR in Big Data systems.

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. L119, 188 (2016).
2. Pham, P, L.: "The applicability of the GDPR to the Internet of Things." Journal of Data Protection Privacy 2.3 (2019): 254-263.

3. Yuan, B., Jiannan, L.: "The Policy Effect of the General Data Protection Regulation (GDPR) on the Digital Public Health Sector in the European Union: An Empirical Investigation." International Journal of Environmental Research and Public Health 16.6 (2019): 1070.
4. Lopes, Isabel, M., Teresa, G., and Pedro, O.: "Improvement of the Applicability of the General Data Protection Regulation in Health Clinics." World Conference on Information Systems and Technologies. Springer, Cham, (2019).
5. Shah, A., Banakar, V., Shastri, S., Wasserman, M., and Chidambaram, V.: "Analyzing the Impact of GDPR on Storage Systems". In 11th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 19) (2019).
6. Gonalves, A., Anacleto, C., and Lus, C.: "An Approach to GDPR Based on Object Role Modeling." World Conference on Information Systems and Technologies. Springer, Cham, (2019).
7. Kotsios, A., Magnani, M., Rossi, L., Shklovski, I., and Vega, D.: An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research. arXiv preprint arXiv:1903.03196. (2019).
8. Camilo, J.: "Blockchain-based consent manager for GDPR compliance." Open Identity Summit 2019 (2019).
9. Krempel, E, and Jrgen B.: "The EU general data protection regulation and its effects on designing assistive environments." Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference. ACM, (2018).
10. Gjermundrd, H., Ioanna, D., and Kyriakos, C.: "privacyTracker: a privacy-by-design GDPR-compliant framework with verifiable data traceability controls." International Conference on Web Engineering. Springer, Cham, (2016).
11. Rantos, K., Drosatos, G., Demertzis, K., Ilioudis, C., Papanikolaou, A., and Kritsas, A.: "ADvoCATE: A Consent Management Platform for Personal Data Processing in the IoT Using Blockchain Technology." International Conference on Security for Information Technology and Communications. Springer, Cham, (2018).
12. Fischer-Hbner, S., Angulo, J., Karegar, F., and Pulls, T.: "Transparency, Privacy and TrustTechnology for Tracking and Controlling My Data Disclosures: Does This Work?." IFIP International Conference on Trust Management. Springer, Cham, (2016).
13. General Data Protection Regulation, https://gdpr-info.eu. Last accessed 20 June 2019.
14. Fernandez-Gago, C., Tountopoulos, V., Fischer-Hbner, S., Alnemr, R., Nuez, D., Angulo, J., Koulouris, T.: "Tools for cloud accountability: A4cloud tutorial." IFIP International Summer School on Privacy and Identity Management. Springer, Cham, (2014).
15. The Absolute Platform, www.absolute.com. Last accessed 20 June 2019.
16. Alien Vault USM, www.alienvault.com. Last accessed 20 June 2019.
17. BigId, https://bigid.com/eu-gdpr/. Last accessed 20 June 2019.
18. BWise GDPR Compliance solution, www.bwise.com/solutions/regulatory-compliance-management/global-data-protection-regulation-gdpr. Last accessed 20 June 2019.
19. Consentua, https://consentua.com. Last accessed 20 June 2019.
20. PrivacyPerfect, https://www.privacyperfect.com/fr. Last accessed 20 June 2019.
21. Hesselman, C., Jansen, J., Wullink, M., Vink, K., Simon, M.: "A privacy framework for DNS big data applications." tech. rep. (2014).
22. Crabtree, A., Lodge, T., Colley, J., Greenhalgh, C., Glover, K., Haddadi, H., Wang, L.: "Building accountability into the Internet of Things: the IoT Databox model." Journal of Reliable Intelligent Environments 4.1 (2018): 39-55.

23. Rhahla, M., Abdellatif, T, Attia, R., and Berrayana, W.: "A GDPR Controller for IoT systems: Application to e-health." WETICE (2019).
24. Ferreira, A., and Joana, M.: "TagUBig-Taming Your Big Data." 2018 International Carnahan Conference on Security Technology (ICCST). IEEE, (2018).
25. Data protection by design and default, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-by-design-and-default/?q=necessary. Last accessed 20 June 2019.
26. Cavoukian, A.: "Privacy by design: The 7 foundational principles." Information and Privacy Commissioner of Ontario, Canada 5 (2009).
27. Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J. H., Metayer, D. L., Tirtea, R., Schiffner, S.: "Privacy and data protection by design-from policy to engineering." arXiv preprint arXiv:1501.03726 (2015).
28. Apache Eagle, https://eagle.apache.org/. Last accessed 20 June 2019.
29. Apache Atlas, https://atlas.apache.org/. Last accessed 20 June 2019.
30. Apache Ranger, https://ranger.apache.org/. Last accessed 20 June 2019.
31. Fischer, H., Simone, Hans, H., and Erik, W.: "Trust and assurance HCI:" Privacy and Identity Management for Life. Springer, Berlin, Heidelberg, (2011). 245-260.
32. Angulo, J., Fischer-Hbner, S., Pulls, T., Wstlund, E.: "Usable transparency with the data track: a tool for visualizing data disclosures." Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. ACM, (2015).
33. Benghabrit, W., Grall, H., Royer, J. C., Sellami, M., Azraoui, M., Elkhiyaoui, K., Bernsmed, K.: "A Cloud Accountability Policy Representation Framework." Closer. (2014).
34. Apache Knox, https://knox.apache.org/. Last accessed 20 June 2019.
35. D'Acquisto, G., Domingo-Ferrer, J., Kikiras, P., Torra, V., de Montjoye, Y. A., and Bourka, A.: "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics." arXiv preprint arXiv:1512.06000 (2015).
36. Wu, Jing, and Zheng, L.: "Research on customer segmentation model by clustering." Proceedings of the 7th international conference on Electronic commerce. ACM, (2005).
37. Cavoukian, A., Chibba, M., Williamson, G., and Ferguson, A.: "The importance of ABAC: attribute-based access control to big data: privacy and context." Privacy and Big Data Institute, Ryerson University, Toronto, Canada (2015).
38. Tokenization, https://www.pcidss.com/listing-category/pci-dss-tokenization/. Last accessed 20 June 2019.
39. Apache Kafka, https://kafka.apache.org/. Last accessed 20 June 2019.
40. Cattell, R.: "Scalable SQL and NoSQL data stores." Acm Sigmod Record 39.4 (2011): 12-27.
41. Yod-Samuel, M., KUNG, A.: " Methods and tools for GDPR compliance through privacy and data protection engineering". In : 2018 IEEE European Symposium on Security and Privacy Workshops (EuroSPW). IEEE, (2018). p. 108-111.
42. Schneider, G.: "Is Privacy by Construction Possible?." International Symposium on Leveraging Applications of Formal Methods. Springer, Cham, (2018).
43. Ferrara, P., and Fausto, S.: "Static Analysis for GDPR Compliance." ITASEC. (2018).