Business Process-Based Legitimacy of Data Access Framework for Enterprise Information Systems Protection

Hind Benfenatki(✉) and Frédérique Biennier

University of Lyon, CNRS, INSA-Lyon, LIRIS UMR 5205, Lyon, France
{hind.benfenatki, frederique.biennier}@insa-lyon.fr

Abstract. Nowadays European context is introducing a new directive for data protection, which imposes new constraints to business owners which manipulate personal data. Among imposed constraints, we find that while a disclosure occurs on user's personal data, the burden of proof is now in the charge of business owners. In this context, data access has to be managed according to what is mentioned in Terms of Service and logged in a way to prove the occurrence of a disclosure or not. This work, part of Personal Information Controller Service project proposes a data-driven privacy control system, based on Collaborative Usage Control (CUCON), allows organizations to manage the access authorizations they provide to stakeholders. The proposed system intervenes in two contexts, which are ad-hoc business processes and while using big data techniques. In fact, new data usage introduces changes in usage-based models since used systems are usually distributed and involving several organizations which can have different definitions for a given role. This framework manages the consistency between already allowed data access rights and potential given rights to a given business stakeholder according to business process's activity affected to him/her. It also warns when a conflict occurs and when the aggregation of the rights granted to a given stakeholder lead to having rights to a sensitive data.

Keywords: Usage-based access control · General data protection regulation
Ad-hoc business process · Big data analytics · Legitimacy of data access

1 Introduction

European Union is about to apply imminently the General Regulation of Data Protection (GDPR). The latter brings changes in data treatment. In fact, among other things it requires from data collectors and processors to explicitly obtain data owners' consent before processing it. GDPR also changes the costs assessment related to the disclosure of a given data. In fact, the damage now concerns not only the provider's reputation, but a compensation is allowed to the data owner which personal data are disclosed. The deployed protection means put in place and their cost intervene in the calculation of the compensation while a disclosure occurs.

GDPR involves several changes in data protection. First, personal data is defined by GDPR [1] as any information relating to an identified or identifiable natural person. It

A. M. Tjoa et al. (Eds.): CONFENIS 2018, LNBIP 327, pp. 146–160, 2018.
https://doi.org/10.1007/978-3-319-99040-8_12

regroups financial data, administrative data, identity data (e.g., name, date of birth), medical data, biometric elements (e.g., fingerprint), connection data (e.g., IP address), localization data (e.g. GPS tracking), activity data (e.g. cookies) and so on.

Second, data usage is changing. In fact, nowadays, enterprises use big data techniques (e.g., data analytics, data mining) on collected data to extract knowledge and generate new data to serve their business. The objective is to examine raw data in order to draw conclusions from this information. These conclusions may represent additional information on users which are not a priori listed in the Data Base Management System (DBMS) of the Information System (IS). Cate et al., have written that *'it is no exaggeration to say that we are nothing more than a collection of data to most of the institutions—and many of the people—with whom we deal'* [2]. *'Our biographies are etched in the ones and zeros we leave behind in daily digital transactions,'* as Stanford Law School Professor Kathleen Sullivan has written [3]. It is now possible to identify, describe and increasingly define us based on collections of zeros and ones [2].

Third, GDPR requires from data consumer to inform data owner about the usage to be done with his/her data. Data usage may have several purposes which are not necessarily communicated to the data owner, thus introducing unawareness of explicit concrete usage to be done with the data. For example, while doing data analytics, the user does not know how its data are analyzed, with which other data his/her data can be merged, and which information can be generated or deduced from his/her data. This leads to hardiness of estimation of identification risks related to such data usage.

Lastly, the development of data analytics shows that Big Data also provide predictive analysis in industry 4.0 which stands for the entry of IT into the manufacturing industry [4]. In fact, old manufacturing processes evolve by integrating a massive amount of sensors. The analysis of sensor-collected data and context data leads to the capability to predict deviations from standard maintenance intervals [4]. Another example of big data usage is healthcare [5]. In fact, a team of Stanford University discovered that the association of two drugs which are Paxil®—the blockbuster antidepressant prescribed to millions of Americans—and Pravachol®—a highly popular cholesterol-reducing drug—generates as side effect the increase of patient's blood glucose to diabetic levels. This discovery has been done by pursuing statistical analysis and data mining techniques in the Adverse Event Reporting System (AERS). The latter is a Food and Drug Administration's (FDA) database which collects adverse drug event reports from clinicians, patients, and drug companies for more than thirty years [5]. Other examples of use of big data can be found in literature [5].

Despite the benefits that can bring latent big data analytics as we saw in the example of drugs side effects, big data analytics may generate and/or manipulate very sensitive data. In big data analytics context, it is the generated data (inferred down from big data) that cause for concern [5]. To deal with that, organizations have to disclose their decisional criteria [5] in order to be able to define access control policies also on generated data. In this context, data protection and user sensitization have to be at the heart of our concerns. Data protection concept encompasses a lot of challenges depending on the environment and context where data are manipulated. In this sense, 'lawmakers around the globe are struggling to find a new balance between the need to protect the information privacy of individuals against the demand to utilize the latent value of data' [6]. In this paper, we are interested in access control in big data analytics.

In this context, how can we manage access control on generated data while dealing with big data analytics?

In the case of data analytics, we are in a context of ad-hoc process due to the unawareness introduced about the analysis process and their results. In fact, ad-hoc processes regroup activities that cannot be predefined [7]. They require a self-organization team since users have to be able to decide what to do and when to do it, and also they must be able to assign work (activities) to other people [7]. In this context, extrapolating CUCON in a semi-structured or ad-hoc business processes context is not easily applicable because it may necessitate the update of role's rights. However, this will imply that all users with the same role are granted while it is about ad-hoc processes. In this context, how can rights be granted using CUCON without upgrading the rights of all users with the same role in ad-hoc processes?

Moreover, as each data may have several allowed usages, this leads to new issues in addition to data access rights, which are: the legitimacy of access and identification risks.

To fit GDPR requirements we focus on new usage related to big data and ad-hoc processes to identify if data access can be granted or not. To this end, we propose a framework for estimating legitimacy of data access in ad-hoc business processes and big data analytics contexts.

This paper is organized as follows. Sections 2 and 3 describe respectively the background describing considered data types' classification and related work concerning data privacy management. Section 4 describes our proposed solution. Sections 5 and 6 respectively discusses our work and draws final conclusions.

2 Background

According to way it is obtained, personal data can be classified in different groups. Explicit data represents the information given explicitly by the user (e.g., while filling out a form).

Collected data is induced by user's activity (for instance, connection activity, or even GPS coordinates). In addition to the use of different personal data, the identification of an individual may also arise from his/her interaction with different services. Then we have an identification from the traces of activity. While these identifications may seem less intrusive because they are not directly related to sensitive information, they nonetheless represent a major risk of privacy violation insofar as cross-linking of data sources can make it possible to link these identities derived from trace data, to a legal identity without the user knowing that he/she is identified.

Generated data results from information construction or data fusion leading to control the way data is used and transformed. In fact, big data algorithms impose to cross many data sources, thus generating information not explicitly given by the user. Furthermore, some providers analyze user's personal data in order to construct an information which is not given explicitly. For example, receiving a travel ticket via Google mail service leads to email analysis so that travel information can be sent and displayed on the user's Android phone and on google calendar.

Deduced data results from human interpretation of at least two data to generate a new information. For example, a school's secretary may deduce from parents' addresses' their marital status, since if the addresses are different, this implies that the parents are separated or divorced.

In this work, we are interested in explicit, implicit, collected, deduced and generated data. In fact, we analyze the right to access explicit, implicit, and deduced data during ad-hoc business processes, and the right to analyze implicit and explicit data as part of data analytics, depending on the potential generated data.

3 Related Work

Traditionally, data protection is mostly achieved by organizing a secured architecture to support the information system deployment as well as access controls (for internal threats). Most of these authorization systems rely on Role-Based Access Control (RBAC) models [8] since it is more suitable to affect the business process access rights to business roles, instead of to business stakeholders. American National Standards Institute (ANSI) defined RBAC [9] as a mechanism which controls access according to users' roles. *'Each role assigns a collection of permissions to users. RBAC assumes that, in most applications, permissions needed for an organization's roles change slowly over time, but users may enter, leave, and change roles rapidly'* [10].

However, new data usage introduces new challenges while using RBAC. In fact, with the advent of cloud computing and big data techniques, data and users are decentralized. On the one hand, in Software as a Service (SaaS) model, data is stored on provider's side or even on subcontractor data storage service. It is accessed by provider's service in order to perform a function and also by service users. In this context, instead of having one organization with different roles to manage, we have three organizations that collaborate, each of them with its proper definition of roles. Furthermore, data access control is managed by service and data storage providers following rules defined by service order. Currently, researchers mostly use encryption to ensure data privacy in the cloud [11, 12].

A Collaborative Usage Control (CUCON) [13] has been designed to manage multiple policies in a context of business federation. A collaboration-context oriented policy model is designed. It enables multiple providers to co-define the policy upon the collaborative work artifact, forming the SSLA. Thus CUCON takes into consideration the attributes of the collaboration context in addition to those regarding the assets, the consumers, and the information system infrastructure. An aggregation algebra is used to 'combine' the individual policies from each 'Stakeholder' in order to ensure policy consistency.

According to Hu et al. [14], current big data architectures are based on distributed processing. They are composed of (1) a Master System (MS) which is responsible for receiving data from big data providers and determining processing steps (task distribution, data distribution, and result collection) in response to a user's request; and (2) Cooperated Systems (CS) which represent trusted systems responsible for big data processing and returns computed results. Hu et al. [14] propose a big data scheme with focus in authorization (access control). Authorization management in big data environment is more challenging than in non-big data, because of the need to synchronize access privileges between the MS and CSs. The authors defined Federated Attribute Definitions (FAD) dictionary which regroups the common attributes' definition used by both of MS and CS to describe their respective access control policies in order to allow a coherent composition of them.

Users privacy may be guaranteed using anonymization of their data. However, anonymization may not be suitable for all circumstances especially in big data processing. In fact, anonymization introduces a high degree of uncontrollability while doing profiling on individual targets anonymized data [15]. To deal with that, pseudonymization represents a good alternative. In fact, it allows to guarantee the accountability of data controller and processor since there will always be a person who can re-identify subjects included in a cluster. It also respects personal data protection obligations since pseudonymization allows to reconstruct the processes of identity masking, by allowing re-identification [15].

Samuel et al. [16] is interested in access control management for healthcare multimedia Big data. The authors proposed a framework for composing and enforcing context-aware disclosure rules for preserving privacy and security. Online user's composed disclosure rules are consistent and are verified for a set of verification properties. The authors illustrated the proposed framework with a healthcare scenario where users own their multimedia patient data pertaining to MRI, X-rays, sonograms etc., and can empower user to control their private data not only in terms of management and access but also allowing the sharing of their data with others whom they authorize, in a private, secure and confidential environment.

Yang et al. [17] propose an access control system for IOT-based healthcare big data to preserve patient's privacy. The access control system proposed is self-adaptive since it allows medical staff to access patient's data in normal and emergency situations. In fact, in a normal situation only the medical staff authorized to access patient's data can access to them. However, in an emergency situation, the patient may not be treated by the same medical staff that is authorized to access its historical medical data. The authors propose that patient's historical medical data can be recovered in emergency application, using a password-based break-glass access mechanism to deal with this dilemma.

These work focus on defining systems and mechanisms to guarantee data privacy using anonymization, encryption, and access control management. Access control management is done by composing protection policies meaning from several stakeholders. However, none of those works is interested in estimating or deducting legitimacy of stakeholder's data access based on his/her existing rights, as it can be the case in ad-hoc business processes and big data analytics.

4 Legitimacy of Data Access Evaluation Framework

Meeting the new constraints from GDPR, especially for data access rights is crucial in nowadays European context. In fact, one of legal obligations of GDPR is to provide strong guarantees that information will not be involved in 'undeclared' usage. Furthermore, in this new directive, the burden of proof is now incumbent on the data collector. The latter must be able to prove that access (even internally) to data has been made in accordance with what has been declared and that the rights do not go beyond what is planned.

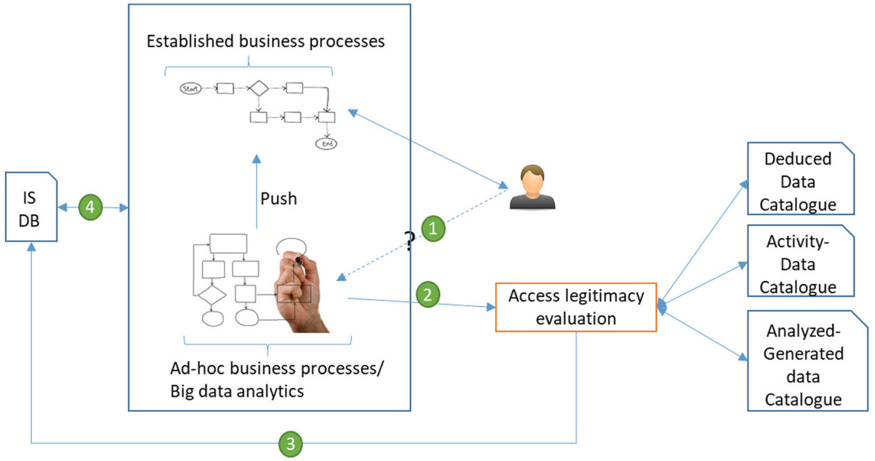
Literature is full of works allowing to address data privacy via access control mechanisms, or anonymization etc. However, there is a lack of access legitimacy evaluation. This work aims to address access legitimacy evaluation in order to manage identification risks associated to each data usage/access and thus to define protection requirement. To this end, we had to:

- Recognize identification risks origins and legitimacy access issues. This has been done by browsing data analytics process. We identified two points:
    - Enterprise Information Systems usually follow business processes. Each task of the business process manipulates data. This supposes the management of access control on each data of each process. Business Processes represent semi-structured data. Tags are associated to them allowing to know the business context of a business process which helps to identify the legitimacy to access some data. If access rights are associated to users instead of activities (business process tasks), this allows to evaluate the business stakeholder's access rights legitimacy instead of the legitimacy of using a given data by a given service. In order to cover both legitimacy evaluations, we used two catalogues correspondences: user-authorized activities, and activity-authorized data.
    - Even if data access management using the previously defined catalogues allows to meet user/service access legitimacy evaluation, identification risks arise. In fact, two activities associated to one user increase the authorized data set leading to increasing identification risks. Indeed, a deduction of an information from two data or more is possible. Furthermore, data analytics using big data technics allows to generate new data from an initial data set. In order to not increase identification risks, we take into account deducted and generated data before granting access rights to a given business stakeholder.
- Determine access legitimacy evaluation mechanism that allows to fulfill the previously cited identification risks and access legitimacy issues.
- Determine evaluation scenarios which highlight the cases of abusive authorizations and inconsistent rejects that can be avoided using our approach.

We present in this section the legitimacy of data access evaluation framework. We first describe the general view and used concepts' definition. We then describe the functioning of the framework.

4.1 General View and Used Concepts' Definition

The architecture of the solution we propose is illustrated in Fig. 1. The proposed framework has as input an access request, and returns a deny/permit access to the request's data. We intervene in two different contexts: ad-hoc business processes and while using big data techniques. The former occurs when an activity is allocated to a given business stakeholder as part of ad-hoc business processes. In this context, the proposed framework allows to match user's potentially allocated activities' data with user's already authorized data. The later occurs when data analytics are done in order to evaluate the access legitimacy of the generated data. It allows to detect conflicts emerging while attributing the task of data analytics to a given business stakeholder according to its authorized data. In other words, while a business stakeholder is involved in a business process, she/he has implicitly data access rights to manipulated data. However, in the context of data analytics, manipulated data generates new ones leading to the growth of access rights.



(1) When a user launches an ad-hoc process, his/her rights are evaluated based on tags used to configure the task (2) to check if this usage is similar to what the user is allowed to do in other business process (step 3 consolidates these information). If the usage is accepted, the legitimacy evaluation is also called each time access to the information system is achieved (4)

Fig. 1. Data access legitimacy evaluation framework's general view

The framework is composed of:

- *Information System. Data Base* regroups structured and unstructured data;
- *Activity-Data catalogue*. Before describing the purpose of this component we first describe the activity concept. An *activity* corresponds to a business process activity. It is described using *tags* describing at a high level of abstraction the function of the activity. For example, in a travel business process, we can have "trip booking" and "agent billing" activities. Each activity may require data to be achieved. For example, a 'trip booking' activity may require to access to client's credentials in order to check if the latter needs a visa for the booked trip, and/or if she/he has the legal age to travel alone. We then describe the association of a business process's activities with the legitimate data to be manipulated;
- *Analyzed-Generated data catalogue*. This catalogue inventories the data that may be generated while doing data analytics in a given set of data. In fact, while requesting big data analytics on a given set of data, one can have the rights to access the requested data but not necessarily the generated ones.

- *Deduced data catalogue*. This catalogue associates to each data set, a data set corresponding to the data that can be deduced by human reasoning (not using data analytics methods). For example, it associates to {mother's address, father's address}→{parent's marital situation}.

4.2 Data Access Legitimacy Evaluation Framework

Request Composition

The user's request is composed of four attributes as follows:

$$\text{Auth\_request} = (\text{Assigner\_ID}, \text{Assignee\_ID}, \text{Activity}, \{\text{Data}\})$$

Where:

- Assigner_ID and Assignee_ID represent respectively who launched the activity and who has to execute it. Assigner and Assignee may refer to the same person;
- Activity, represents the action that has to be performed by the assignee. It is described using tags. Activity has "data analytics" value while doing big data analytics;
- Data: depending on the request context (ad-hoc business processes, big data analytics), Data attribute may refer differently. In ad-hoc business processes data access request, data represents the list of the data the user wants to access, while in data analytics request, Data attribute concerns the data the user will manipulate as part of data analytics. In the first case, Data attribute may be facultative since a catalogue inventory lists the correspondences between an activity and the data it manipulates (Fig. 2).

Steps

The data access legitimacy evaluation process is illustrated in Fig. 2. It follows the steps described below, depending on the request context. We consider two cases: big data analytics and other ad-hoc processes. The latter is defined in this work as a business process for which at least one activity does not figure among considered ones. In both cases, we check if the assignee's role has the rights to access the requested/generated data. The verification process is described in the section below. If yes, the authorization is accorded. Else if the assigner is a "*super user*", the latter has two choices as follows:

(1) Granting user's role with the access to the requested data,

(2) Fine-graining user's role. This means that a fine-grained role may be created and allocated to the user to which the data access has to be granted. This new role inherits access rights of the generic role, i.e., the role previously played by the assignee. Before validating the fine-graining process, the possible deduced data from the ones the user is now authorized to access are calculated. The estimation of deduced data is done from the *"deduced data catalogue"* (cf. Fig. 1). The objective of fine-graining is here twofold: it respects the original access grants, and it avoids to generalize data access rights to people with the same role but

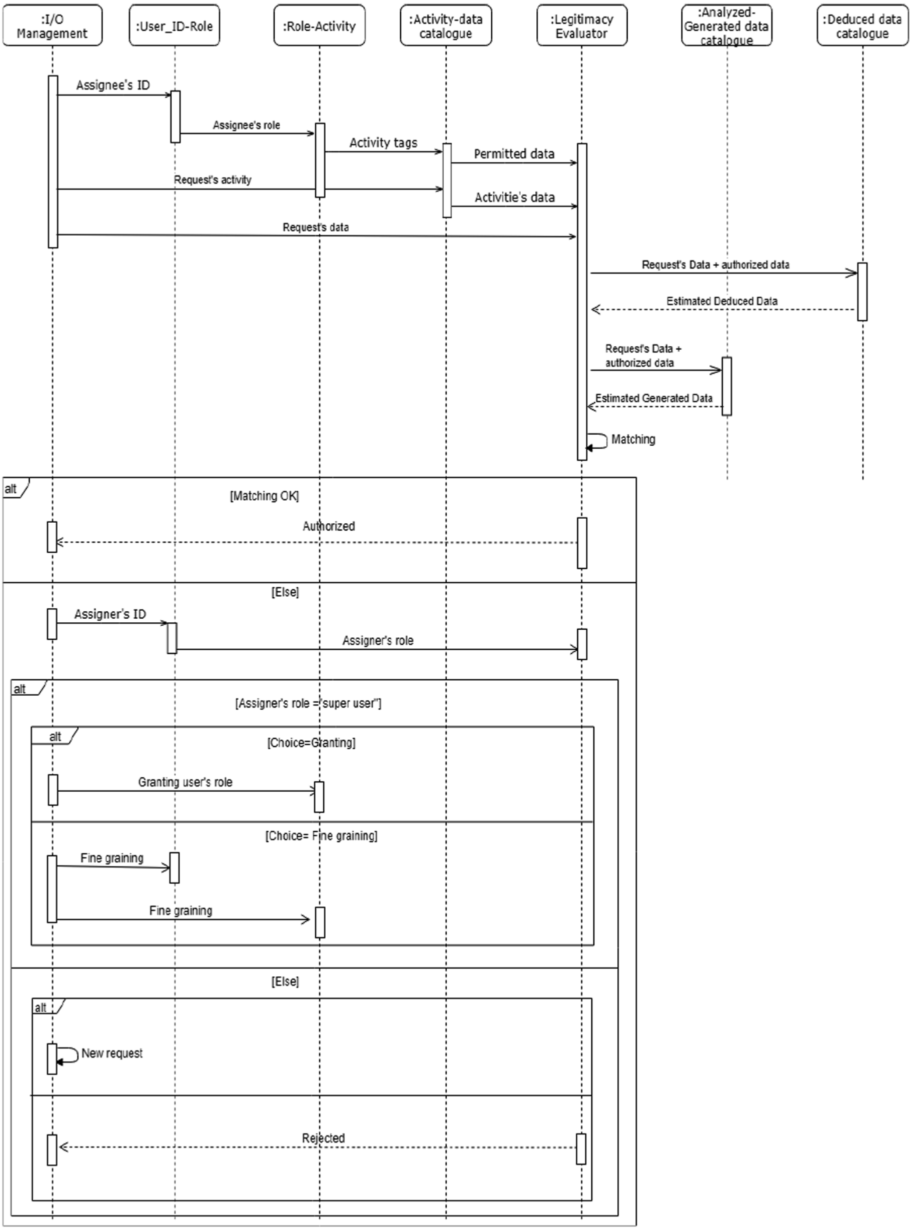


Fig. 2. Data access legitimacy evaluation's sequence diagram

which do not necessarily require the granted access rights. Fine-graining can be used to avoid the deduced/generated data by defining different levels for each role. Each of them access to a part of the data that allow the generation of unauthorized ones.

Let us take the example of an organization that would like to grant the right to users with *"secretary"* role to access to employees' addresses. However, the organization counts 5 secretaries. To grant the access right to only one of the 5 secretaries, we fine-grain the role secretary by creating a type *"super-secretary"* for the role *"secretary"* which inherits all secretary's rights to which is added the access right to employee's addresses. This leads to the creation of a hierarchy in roles' definition.

In order to meet coherence while defining access to each role, we analyze the impact of access rights granted to each role by considering the possible deduced data from the originally granted. If the potentially deduces data do not figure in the authorized ones, access rights have to be revised. The revision may concern the upgrade/downgrade or fine-graining of access rights granted.

Else if the matching is not verified, and the assigner has not a *"super user"* role, two choices are possible: changing the request's assignee in order to find a user who fits required access rights, or the authorization is rejected.

Verification Process

We proceed to assignee profiling to determine its permitted data based on its ID following the stages below:

- **Assignee_ID-Role correspondence:** in this work, the rights to access data are allocated depending on business processes the user is involved in, which in turn depends on user's role. In this sense, while a user requests an access for a given activity (which involves the manipulation of given data), we have firstly to check the role of the user. This stage allows to reply to that question;
- **Role-Activity correspondence:** this second stage is concerned with the identification of the tags of the business processes' activities affected to the role played by the assignee.

(1) If the requested activity is not a "big data analytics" one, and figures on authorized activities for the role concerned by the request, the authorization is accepted.

(2) If the requested activity is not a "big data analytics" one, and does not figure on authorized activities for the role concerned by the request, we proceed to **activity-permitted data correspondence**. Each business process manipulates data. The objective of this step is twofold: deducting assignee's authorized data and recovering the manipulated data for the requested authorization's activity. If assignee's authorized data covers requested activity's data, then the request is authorized, and the assignee's role is updated with the corresponding activity. Permitted data for a given activity is considered as explicit. However, some data can be deduced from explicitly authorized ones. For this reason, in addition to considering explicitly authorized data access, we consider also the possible deductible data as authorized data since the user has access to that information anyway.

(3) If the requested activity is a "big data analytics" one, a verification is done regarding which kind of data the assignee has the right to analyze. To do that we **estimate potential generated data**. The objective of this stage is to highlight potential inconsistency between the data that can be generated while doing data analytics on a given set of data by a given business stakeholder, and the data that are really permitted to that user. The data that can be generated while doing data analytics is estimated based on the construction of a knowledge base of generated data from a given data set. For example, while analyzing the transport flow of a person, we can generate his/her work and home neighborhoods. The base of knowledge is enriched every time a data analytics are done, and using learning techniques.

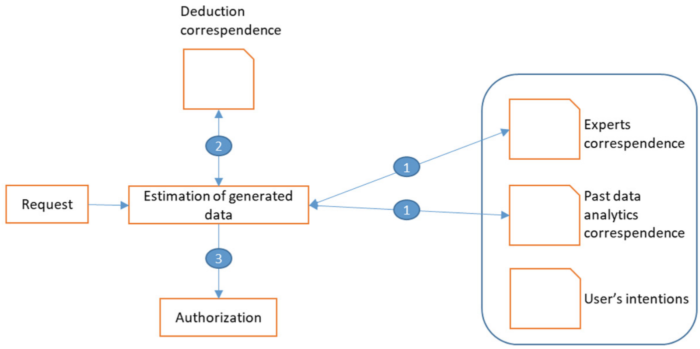
The estimation of generated data is illustrated in Fig. 3. The information is deduced from three sources which are:

- Past data analytics-based correspondence, regroups past correspondences between analyzed data and generated ones. This correspondence is updated every time a new data analytics is done. The update is done when the correspondence for a given set of data does not figure for the running data analytics, the latter is added to the correspondence table, and when the correspondence exists for the running data analytics but the generated data do not match. In the last case, the update consists of updating the generated data in order to include the generated data that do not appear.
- Experts-based correspondence, regroups the correspondence between analyzed data and generated ones described by experts. This represents the reference correspondence.
- Assignee's intention-based correspondence represents the correspondence between the data to be analyzed and the expected generated data for the user.

We rely on these three correspondences in a way to enrich the knowledge base. The estimation of generated data follows the correspondences' priority of trust as follows: experts correspondence → past data analytics correspondence → assignee's intention correspondence.

Following the estimation process described in Fig. 3, if the user has defined his/her intention regarding the data analytics, i.e., the data set which she/he is expecting by analyzing the request's data, a request's validation is done by comparing user's intention from his request, with the correspondence tables. If the data the user wants to analyze allows to generate intended data, the request is considered as valid and we proceed to the next phase, else the estimation tool returns an error. If the user has not defined his/her intention regarding the data analytics, generated data are estimated following request's data. This is done from the correspondence tables following the priority trust defined above.

- **Data matching:** this occurs by comparing syntactically authorized data of the assignee business stakeholder from profiling phase with the desired data/possible generated data. If the desired data and generated data figure in the list of the legitimate data of the user, then the request receives an authorization.



(1) The estimation tool requests for the potential generated data from request's data to experts and past data analytics correspondences
(2) Request's validation
(3) Authorization result

Fig. 3. Estimation of generated data

5 Evaluation and Discussion

To evaluate our work, we compare the proposed framework with RBAC solution. Table 1 illustrates the evaluation of our framework according to the following criteria:

- Request validation evaluation;
- Kind of data covered;
- Request types. We use this criterion to evaluate if a distinction is done between request types (activity types) or if all activities are evaluated at the same level, which introduces abusive authorizations;
- Resolution of unknown access to activities/data criterion is used to evaluate the presence of inconsistent rejects;
- Fine-graining while upgrading role's access rights;
- Estimation of generated data while evaluating a data analytics request;
- Estimation of deductible data while evaluating a request.

The objective of this work is to allow role-based access control to data in a context of ad-hoc business processes and while doing big data analytics. This choice is guided by the fact that in both cases, the access authorizations may not be predefined to a given role because ad-hoc business process are constructed as time goes by on the one hand, and because big data analytics introduces an unknown which is the possible generated data on the other hand.

It appears from the table that the advantage of our framework compared to classical RBAC lies in the fact that it distinguishes data analytics from other activities. This allows to take into account generated data while doing data analytics, thus considering the access authorization for the real accessed data, i.e., the analyzed data and the generated ones. This is not allowed in traditional RBAC.

Table 1. Comparison between classical RBAC and proposed CUCON-based legacy framework

Evaluation criteria	Our framework	RBAC
Request validation	Every time a new authorization of type "data analytics" is introduced for a given business role	NC[a]
Data types	- Explicitly permitted explicit and implicit and - Generated data - Deductible data	No visibility on permitted generated data and deductible data
Request types	- Data analytics - Others	Does not distinguish data analytics requests from the others. This leads to the same request treatment
Resolution of unknown access to activities/data	Is done using deductible data	NC
Fine-graining	Is done to avoid granting rights to all role's users while granting role's access rights	NC
Estimation of potential generated data	Is done while evaluating a data analytics requests	NC
Estimation of deductible data	Is done while evaluating all requests	NC

[a] Not Considered

It also allows to resolve unknown activities/data access requests. In fact, when an assignee requests for the access to an activity/data which is not listed among its authorized ones, instead of rejecting the requests from the outset, an evaluation of the data the user can deduct from its authorized data are estimated and added to the set of user's authorized data.

Table 2 illustrates the authorization results for two requests assuming that:

– The user has authorized access to data {D1, D2},
– One can deduce D3 from {D1, D2},
– The data analysis of {D1, D2} leads to generate {D3, D4, D5},

Out of Table 2, we observe that for both requests we have contradicted authorization responses depending on the used method. In fact, classical RBAC does not authorize the read access to D3 while the later can be deduced from {D1, D2} to which the user already has authorized access. This constitutes an inconsistent reject. Furthermore, classical RBAC authorizes the analyze (data analytics) access of {D1, D2} while the later can generate {D3, D4, D5} to which the user does not have authorized access. This constitutes an abusive authorization. Our solution allows to avoid both inconsistent rejects and abusive authorizations.

Table 2. Comparison between our solution and classical RBAC illustrated with an example

Request	Our framework	RBAC
Read (D3)	Yes	No
Analyze (D1, D2)	Non	Yes

6 Conclusion

This work is part of a project that investigates strategies for preserving data protection, in particular, while granting data access. Our approach evaluates the legitimacy of data access and manipulation based on user's role. In fact, we use a Role-Based Access Control mechanism. The legitimacy of data access intervenes in two modes: while a user requests for an activity in the context of ad-hoc business processes, and while a user requests for a data analytics grant based on a given data set. In fact, the need for the latter is all the more sensitive because data analytics generate data which are more informative than the manipulated ones. In this paper, we presented the conceptual approach proposed to manage the legitimacy of data access and manipulation.

As part of our future work, we are planning to allow a more restrictive access while possible. For example, when a tax employee needs to access gross taxable of insurance potential client's, we extract the information from a pay slip, instead of giving the pay slip to the user, since it provides more information than the needed one.

Acknowledgments. This work is partly supported by the Personal Information Controler Service (PICS) project, co-sponsored by the French Secrétariat Général pour l'Investissement, the French Direction Générale des Entreprises and Bpifrance under the "Investissement d'avenir - Protection des données personnelles" grant.

References

1. GDPR, Art.4. [En ligne]. http://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm
2. Cate, F.H., Kuner, C., Millard, C., Svantesson, D.J.B.: The Challenge of "Big Data" for Data Protection (2012). chez Articles by Maurer Faculty. 2620
3. Sullivan, K.M.: Under a watchful eye: incursions on personal privacy. In: The War on Our Freedoms: Civil Liberties in an Age of Terrorism, New York (2003)
4. Becker, T.: Big data usage. In: Cavanillas, J., Curry, E., Wahlster, W. (eds.) New Horizons for a Data-Driven Economy, pp. 143–165. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-21569-3_8
5. Tene, O., Polonetsky, J.: Big data for all: privacy and user control in the age of analytics. Northwestern J. Technol. Intell. Property **11**(5) (2013)
6. Padova, Y., Schönberger, V.M.: Regime Change? Enabling Big Data through Europe's New Data Protection Regulation. Columbia Science and Technology Law Review, pp. 317–339 (2016)

7. Gromoff, A., Stavenko, Y., Evina, K., Kazantsev, N.: Ad-hoc business process management in enterprises as expert communities. In: Chiu, D.K.W., Wang, M., Popescu, E., Li, Q., Lau, R. (eds.) ICWL 2012. LNCS, vol. 7697, pp. 251–260. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-43454-3_26
8. Sandhu, R.S., Coyne, E.J., Feinstein, H.L., Youman, C.E.: Role-based access control models. IEEE Comput. **29**(2), 38–47 (1996)
9. American National Standard for Information Technology: Role Based Access Control, ANSI INCITS 359-2004 (2004)
10. Kuhn, D.R., Coyne, E.J., Weil, T.R.: Adding attributes to role-based access control. IEEE Comput. **43**(16), 79–81 (2010)
11. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan, S.U.: The rise of "big data" on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
12. Zhou, L., Varadharajan, V., Hitchens, M.: Enforcing role-based access control for secure data storage in the cloud. Comput. J. **54**(10), 1675–1687 (2011)
13. Su, Z.: Applying Digital Rights Management to Corporate Information Systems, Lyon (2012)
14. Hu, V.C., Grance, T., Ferraiolo, D.F., Kuhn, D.R.: An access control scheme for big data processing. In: International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom) (2014)
15. Bolognini, L., Bistolfi, C.: Pseudonymization and impacts of Big (personal/anonymous) Data processing in the transition from the Directive 95/46/EC to the new EU General Data Protection Regulation. Comput. Law Secur. Rev. **33**(2), 171–181 (2017)
16. Samuel, A., Sarfraz, M.I., Haseeb, H., Basalamah, S., Ghafoor, A.: A framework for composition and enforcement of privacy-aware and context-driven authorization mechanism for multimedia big data. IEEE Trans. Multimedia **17**(9), 1484–1494 (2015)
17. Yang, Y., Zheng, X., Guo, W., Liu, X., Chang, V.: Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. Inf. Sci. **21**(8), 1–26 (2018)