SN Computer Science (2020) 1:210
https://doi.org/10.1007/s42979-020-00215-x

ORIGINAL RESEARCH

On Enabling GDPR Compliance in Business Processes Through Data-Driven Solutions

Rashid Zaman[1] · **Marwan Hassani**[1]

Received: 20 April 2020 / Accepted: 3 June 2020 / Published online: 21 June 2020

Abstract
The collection and long-term retention of excessive data enables organisations to process data for insights in non-primary processes. The discovery of insights is promoted to be useful both for organisations and the customers. However, long-term possession of data on one hand risks the privacy of data belonging beings in cases of data breaches and on the other hand results in the customers distrust. General Data Protection Regulation (GDPR) abstractly defined the data processing boundaries of the personal data of European Union's citizens. The *processing principles* of GDPR, in line with the spirit of *privacy by design and default*, provide directions on the collection, storage, and processing of personal data. Concomitantly, the data subject rights provide customers with necessary control over their personal data stationed at the data controller's premises. The *accountability* principle of GDPR requires compliance in place and also the ability to demonstrate it. In this work, we are providing three solutions to enable GDPR compliance in business processes. First, we are proposing intra-process data degradation, a solution for continuous data minimisation during the course of business processes. The proposed approach results in reduced data maintenance and breach losses. Second, we adapt process mining techniques for ascertaining compliance of business process execution to data subject rights. Finally, we present a scheme to utilise differential privacy technique to enable GDPR-compliant business process discovery. Additionally, we offer links to two effective tools that demonstrate our first and second contributions.

Keywords GDPR · Business processes · Process mining · GDPR-Compliance · Data minimisation · Differential privacy

Introduction

Business processes collect, generate, or manipulate data of related entities, the beings to whom the process is related and the organisational resources related to the process. The goal of obtaining "data-driven" business models motivated organisations to collect and store as much data as possible and to process it even for non-primary purposes in order to optimise and enhance the organisational processes and maximise the business gains. To cope with the situation, the European Union (EU) introduced the General Data Protection Regulation (GDPR) which abstractly covers all the aspects of the data lifecycle. On one hand *data processing principles* delimit the collection, processing, storage and archiving of personal data to the necessary extent, and on the other hand data subjects are empowered by granting rights over their personal data.

Enabling end-to-end GDPR-compliance in business processes is nontrivial [31]. As forward compliance, processes may need to be completely redesigned or partially overhauled to inculcate the applicable GDPR provisions, especially the data processing principles and data subject rights. GDPR's *accountability* principle requires the compliance to be demonstrable as well. Therefore, once the believed-to-be GDPR-compliant business process is implemented, the execution of the process needs to be monitored. As backward compliance, the execution data shall also be analysed for ascertaining compliance with the applicable GDPR provisions, and detect deviations, if any.

This article is part of the topical collection "Privacy, Data Protection and Digital Identity" guest edited by Fernando Boavida, Andrea Praitano and Georgios V. Lioudakis.

✉ Marwan Hassani
m.hassani@tue.nl

Rashid Zaman
r.zaman@tue.nl

[1] Process Analytics Group, Faculty of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

In this work, we are providing three solutions that contribute towards enabling GDPR-compliance in business processes. To be particular, our three solutions are contributing towards enabling (1) compliance to data minimisation, (2) compliance to data subject rights, and (3) compliance to data security and pseudonymisation. Additionally, we offer links to two tools that we contributed to demonstrate the effectiveness of our first and second solutions. In the following paragraphs, we are introducing these three solutions.

Data minimisation is one important pillar of data processing principles and enabler of the *Privacy by design and default*, a requirement by Article 25 of the GDPR. Data minimisation requires that personal data shall be collected commensurate with the legitimate processing purpose(s). Data minimisation, essentially taking into consideration data-minimality at collection stage, inherently minimises the exposure aftermath. Data degradation concept advocates incremental data minimisation in chained processes, where the same data is processed for multiple purposes with varying granularity. In essence, data shall be *irreversibly* degraded at specific instances in-between these chained processes to the extent that the resulting precision is sufficient for the successful completion of following process(es) in the chain. The term *precision* is used in the context of granularity and information level of data elements. In this article, we are taking data degradation one step further by proposing *intra-process* data degradation, where data shall be continuously and irreversibly degraded during the course of an individual process. In many types of processes, the activities at a later stage of the process might either require specific data elements at a less *precision* than that required for the initial stage activities, or not require this data at all. These data elements become, precision-wise, fully or partially superfluous during the execution of such a process. Therefore, these data elements can be degraded through precision and retained information reduction techniques such that the relevant process can still be successfully completed. A degradation policy is the guiding force for such data degradation. We are presenting novel intra-process data degradation policies.

Process mining techniques [1] provide insights on business processes relying mainly on process execution data. Process discovery techniques discover the de facto process model out of event data. Conformance checking techniques check for the harmony between a perceived business process model and the execution of the process in business environment. Compliance checking techniques, a variant of conformance checking, align event logs with business rules and behavioural constraints in order to detect case-level deviations, if any. The diagnostic information may be used to discover the factors leading to deviations and devise evidence-based corrective measures. We are adapting compliance checking techniques for ascertaining compliance to certain GDPR provisions, mainly related to the data subject rights.

Almost all the process mining techniques process event data, which may contain sensitive data of the data subjects and the related organisational resources. Therefore, these techniques shall also adhere to the data security and data pseudonymisation requirements of the GDPR. Differential privacy is a statistical technique which reveals noise-added information about groups of subjects in a dataset without revealing information about individual subjects. The technique has been found effective and privacy-preserving in many domains where sensitive data is processed. We are presenting our vision on making process discovery techniques privacy-preserving, in line with the GDPR data security requirements, by utilising differential privacy technique.

The remainder of this article is structured as follows. Section 2 provides an overview of the related work on data minimisation, data degradation, process mining techniques, and data privacy in process mining. Section 3 details our intra-process data degradation approach, different data degradation policies, and a proof of concept implementation to demonstrate its efficacy. In Sect. 4, we tackle the GDPR compliance in business processes from two different perspectives. First, we discuss post-execution GDPR compliance of business processes with respect to data subject rights. Process mining techniques tailored for the purpose are evaluated. Second, we present our vision and a differential-privacy-based envisaged setup for enabling privacy-preservation in process discovery. Finally, Sect. 5 concludes this article with a discussion of the major relevant challenges we foresee and an outlook on the future work.

Related Work

Domain and use case specific data minimisation at the data collection stage has been addressed through different approaches in the works of [2–5]. Hilderman et al. in [6] introduced *domain generalisation graphs*, which laid the groundwork for data degradation. Primarily aimed at large databases, the different hierarchical levels of the domain generalisation graphs are utilised for rolling up and drilling down the data in the database. The authors in [7, 8] build upon the domain generalisation graphs of [6] to propose degrading data elements by retaining only the information in the graph levels required for future processes. For this purpose, the authors utilise the hardwired life-cycle policy model of [9]. The work in [10] attempts to personalise the time-based life-cycle policies to stake-holders having varying preferences. In [11], the authors demonstrate the viability of their approach by deploying data degradation for enhancing privacy in ambient intelligence.

Our first contribution on data minimisation sounds similar to [7–9], but is different on three grounds. First, the approach proposed in [7–9] is time-based recurrent and inter-process, where multiple processes having distinct and mutually exclusive boundaries use the same data. Precision of data elements remains the same during the course of an individual process. Our proposed approach is intra-process and has three different data degradation policies. Second, the life-cycle policy model of [9] utilised in [7] and [8] is database-centric while our data degradation policies are more process-centric. Third, the approach in [7–9] being database-centric, highlights the data degradation implementation related challenges therein. Our process-centric approach highlights data degradation implementation related challenges being faced from business processes perspective. A theoretical approach that applies data degradation but in a self-triggered way has been proposed in [12], where data become unusable after a designated period. The 'Sticky policies' approach of [13] proposes appending allowed usage and associated obligations as machine-readable policies to date transferred outside organisational boundaries.

A considerable number of process discovery techniques have been devised in the past two decades. However, only the family of inductive miners [14] guarantees to discover sound process models [1]. Business process models may evolve over time due to multiple reasons like function creep or concept drift. Therefore, we need sophisticated and efficient process discovery techniques that are able to deal with streams of events [15, 16, 36]. Alignments-based conformance checking techniques [17, 18] are considered as standard. The compliance checking techniques in [19–21] are able to deal with business rules but only when those are formalised as Petri nets.

The maintainability of event log's privacy in process mining tasks has been researched in [22] and [25]. In [22], personal data is *k*-anonymized [23, 24] before initiating the discovery process. The authors in [25] use a differential privacy agent [32–35] to provide noisy statistics on process traces to an untrusted process mining technique thereby preserving privacy of information contained in the event log. The query-synthesis mechanism and therefore the utility of the resulting discovered process model are improvable.

In [26], the authors analyse the extent to which transparency in right of access is possible to be achieved in the case of police and other law enforcement agencies. Our solution on enabling compliance to data subject rights is advancing our foundational work in [29, 30] towards more concrete framework implementations using ProM Framework [1] in the context of the BPR4GDPR[1] EU H2020 project [31]. However, it differs in domain and scope from [26] as we are considering business processes and technical considerations in implementation of GDPR provisions therein. The compliance assessment framework of [27] is used for assessing the compliance of actions of a user at run-time by asking questions, in contrast to our compliance checking at post-execution stage.

Enabling Compliance to Data Minimisation

In this section, first we discuss on the GDPR article(s) relevant to data minimisation. Next, we present our proposed tools and techniques, and methodology for compliance with the discussed GDPR articles, along with presenting the important intra-process data degradation policies. Finally, we demonstrate the efficacy of our approach with a proof of concept implementation.

Relevant GDPR Articles

Article 5 of the GDPR articulates six important data processing principles, namely: (1) lawfulness, fairness and transparency, (2) purpose limitation, (3) data minimisation, (4) accuracy, (5) storage limitation, and (6) integrity and confidentiality. Together these principles attempt to delimit the collection, storage, and processing of data. Adaptation of these principles in spirit may necessitate thoroughly reworking business processes. In this work, we are focusing only on data minimisation principle. This principle requires personal data to be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".

Article 25 related to *privacy by design and default* requires data controllers to "implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects". Data minimisation has been addressed in Article 47 to be necessarily specified as part of binding corporate rules. Article 89 adds data minimisation into the list of safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

In the literature, data minimisation is portrayed as a criteria to be taken care of at the data collection stage. We advocate the consideration of data minimisation as a continuous function throughout the process life, thereby reducing the impact in case of a data breach. Our data minimisation approach should not be confused with data access minimisation. In such access management solutions, the data remain unchanged but only the access is controlled such that

[1] https://www.bpr4gdpr.eu/

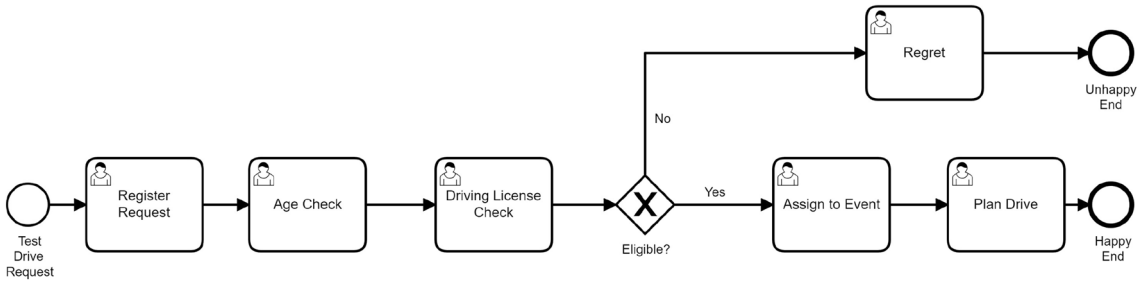


Fig. 1 Car dealership test drive process model

different roles have access to different levels or subsets of the same data.

Compliance Tools and Techniques

We present an example business process to be referenced throughout this section. Then, we define our proposed intra-process data degradation approach, followed by introducing the necessary building blocks of the approach.

Running Use Case Example

We consider a simple test drive process of car dealerships (cf. Fig. 1). According to the process specification, car dealerships arrange test drive events twice a year: first-quarter and second-quarter of each year. Customers interested in test driving a latest car in one of the up-coming test drive events can register in the car dealership information system through an online portal. Personal data like date of birth, driving license number, contact number, mailing address and many other fields are provided for the registration to be successfully completed. Date of birth, driving license number and some other fields like occupation are required for eligibility assessment of test drive candidates. After performing this eligibility assessment, only eligible candidates are assigned to one of the test drive events. Later on, the candidates are contacted at some time prior to the planned test drive event for scheduling a test drive. The duration between registering the test drive request and the actual test drive taking place may span over several months.

Definition 1 (Intra-process Data Degradation) Personal data should always be degraded at suitable points in the process(es) timeline such that the resulting precision and retained information remain sufficient for the successful completion of the process(es), in line with the legitimate processing purpose(s).

Our intra-process data degradation definition highlights two important aspects for data degradation in business processes, namely: *suitable points* and *precision and retained information*, and relatedly the techniques to alter precision and retained information of data elements.

Precision and Retained Information Reduction Techniques

Data elements provide information required for the execution of the activities in business processes. An absolute unit for measuring the information contained in the data is hard to realise, however sufficiency for the processing purpose is an acceptable indicator in the context of business processes. Depending on the purpose of the activities, the same data element with different levels of contained information may be sufficient for the execution of different activities. A perfect knob for tuning the level of contained information of data elements does not exist. Thus, by changing the granularity and the precision of the data element values, it is possible to change the level of contained information or in cases introduce ambiguity to the contained information. The possibility of information reduction provides basis for our intra-process data degradation.

In this work, we are considering two information reduction techniques: vertical and horizontal. Domain generalisation graphs of [6] provide means for reducing information of data elements in vertical fashion. The information contained in a data element may be assigned to different hierarchical levels of a domain generalisation graph such that the level of information at leaf node is very specific and that at root node is very generic. Depending on the context, discarding the information at lower levels reduces the contained information.

Figure 2 presents the domain generalisation graph of a house address in The Netherlands. The root node precisely locates a house while the node one level up makes up into the street address. Street level abstracts to the level of administrative division, divisions one level up combine into cities, which abstracts to the level of a province. Provinces ultimately unite into the country. The precision of a complete address element, consisting of values for all the nodes, can be therefore reduced by traversing up the levels of the address domain generalisation graph and dropping the information contained in the lower levels.

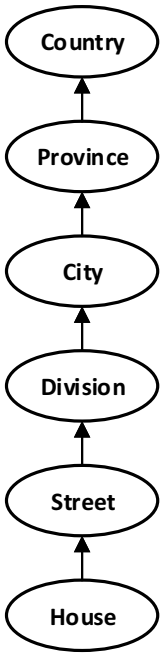


Fig. 2 Domain generalisation graph of a house address in The Netherlands

Horizontal information reduction techniques transform a data element into some alternate representation such that the contained information is reduced or in other words the entropy is increased. For instance, consider `date of birth` data element of a person. A complete date of birth is usually represented in *dd-mm-yyyy* format. As per the requirements of the context, the usual date of birth representation can be transformed to {child, teen, adult, man, old} configuration. In other cases, trimming the day, month, year or a suitable tuple of the `date of birth` data element reduces the contained information and therefore the traceability value of the data element. Similarly, data element `gender` having value `male` can be transformed to ¬ `female` when the initial specificity is no longer required.

Data Degradation Policies

Definition 2 (Data Degradation Policy) A data degradation policy defines the recurrence criteria for initiation of data degradation during the course of a business process.

A data degradation policy is the driving force of our intra-process data degradation approach. It caters the *suitable points* aspect of our intra-process data degradation definition. In this article, three of such policies are proposed: time-driven, event-driven, and event-driven with margin. Each data degradation policy is intrinsically suitable for specific variants of business processes.

Definition 3 (Time-driven Data Degradation Policy) In this policy, the values of sensitive data elements shall be degraded to the next possible information level at defined time intervals $\{T_i, T_j, T_k, \ldots\}$ in the process timeline.

Usually, the degradation time instances are equi-distant, i.e. $\Delta T_{ij} = \Delta T_{jk}$ for all *i, j, k* values, where *i, j, k,* ⋯ are alternating data degradation instances. Referring to the example business process of Fig. 1 and the policy depicted in Fig. 3a, a typical time-driven data degradation policy would be *"degrade data element* `date of birth` *one level every week"*.

The periodicity of data degradation instances, or the *suitable points* aspect of our intra-process data degradation, depends on multiple factors such as data degradation costs, completion time of the process activities, usual process duration, processing nature of the process, i.e. single-instance or multi-instance. Single-instance processes usually batch process data related to multiple cases such as the customer segmentation process for instance. While, in multi-instance processes, an individual instance of the business process is initiated for each individual case such as a bank loan application process for example.

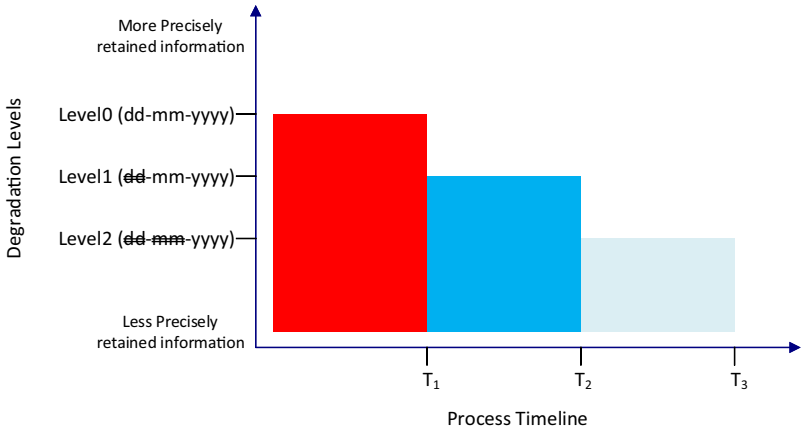
From a data repository perspective, the time-driven data degradation policy is suitable for the environments where the data degradation costs are high. The unavailability of the repository during data degradation operations is also an important factor to be considered here. This policy is recommended in cases where the unavailability of the repository highly affects the performance of the core business process. The composition of the data repositories is a further critical factor to be taken into account. If the process data is scattered or replicated over redundant (heterogeneous) repositories then the infrequent time-driven data degradation policy is adequate.

For the time-driven data degradation policy to be effective and starvation-safe, activities in the business process should have time-bounds on completion. Business processes, however, are prone to deviations from normal behaviour in both the control and time perspectives. Therefore, the time-driven data degradation policy can potentially lead to data starvation in case of deviations where activities are completed later than expected. Additionally, this policy degrades data at pre-defined time instances and as such data can be unnecessarily retained although it may have become superfluous. Therefore, this policy is considered suboptimal.

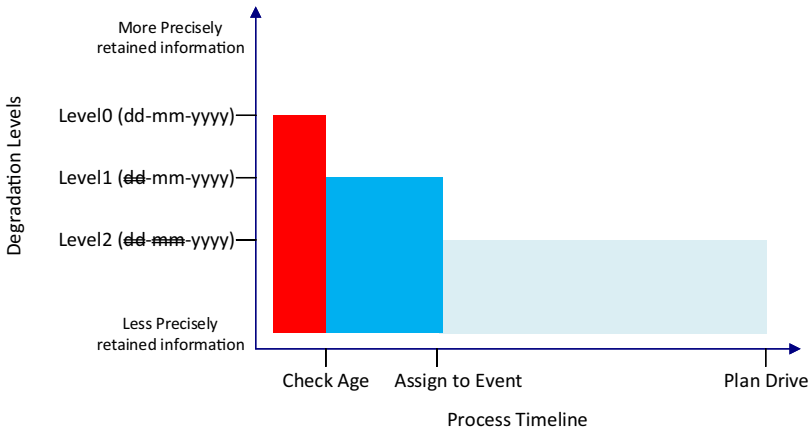
The event-driven data degradation policy caters for the shortcomings of the time-driven data degradation policy.

Definition 4 (Event-driven Data Degradation Policy) In this policy, the values of sensitive data elements shall be degraded to the next possible information level as soon as the current information level is no longer required for successful execution/completion of the process, or in other

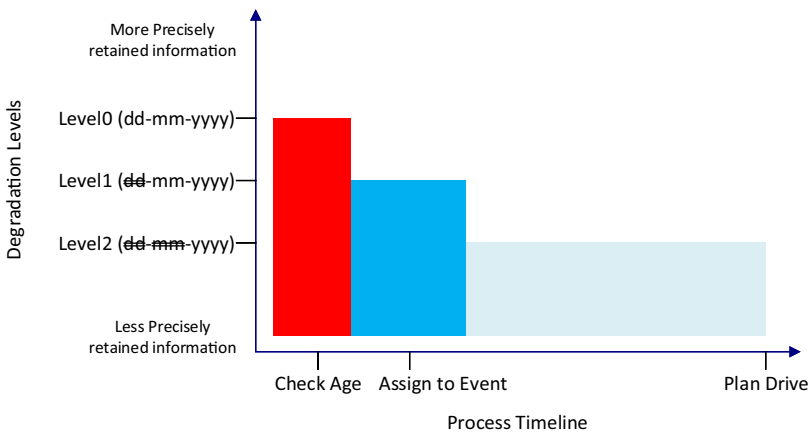
Fig. 3 Data degradation policies discussed in this work for data element `date of birth` in the context of example process of Fig. 1



(a) Time-Driven Data Degradation Policy.



(b) Event-Driven Data Degradation Policy (cf. Figure 1 for the events in the process model).



(c) Event-Driven Data Degradation with Margin Policy (cf. Figure 1 for the events in the process model)

words soon after the last process activity requiring the data at the current information level is executed.

Data degradation initiation instances, or the *suitable points* aspect of our intra-process data degradation, are linked to the completion of specific process activities. The completion of these activities marks the current information level of a specific data element to be superfluous and renders it eligible for degradation. Referring to our example process of Fig. 1 and the event-driven data degradation policy in Fig. 3b, the value of data element `date of birth` is degraded one level as soon as the activity *Check Age* is executed and later degraded one further level when the activity *Assign to Event* is executed.

From the perspective of the data repository, the event-driven data degradation policy is suitable for the environments where data degradation costs are low and the unavailability of the repository during data degradation operations has a negligible effect on the performance of the core business process. From the business process perspective, the event-driven data degradation policy is suitable for both single-instance and multi-instance categories of business processes. Business processes with evenly or sparsely distributed activities constitute an ideal case for the event-driven data degradation policy.

The control-flow deviations are a major threat to the event-driven data degradation policy. Activities triggering data degradation can be executed ahead of their predecessor activities which consequently leaves the latter starving for data at pre-degradation information level. In case of deviations with respect to the time perspective, processes with the event-driven data degradation in place are unlikely to suffer from data starvation. In the worst case, this type of deviations can result in performance bottlenecks where multiple process instances get synchronised which results with a spike in the data degradation loads.

Definition 5 (Event-driven Data Degradation Policy with Margin) In this policy, the values of sensitive data elements shall be degraded to the next possible information level at suitable offsets/margins after the last process activity requiring the data at the current information level is executed.

This policy is suitable for category of processes where the data should obligatorily be retained for a specified period of time, either in accordance with an applicable regulation or some sort of preemption is expected. Different potential scenarios could be: (1) audit teams can randomly pick cases to be analysed for compliance and therefore data should be retained for some specified period, (2) customers are entitled to object to an automated decision within a specified time frame, or (3) the management occasionally inspects cases leading towards an undesirable result such as loan applications leading towards rejection.

In all the previously mentioned and many more related scenarios, data should not be degraded promptly but should instead be retained for a specific period of time, even if it is believed to be superfluous. Referring to our example process of Fig. 1 and the event-driven data degradation with margin policy of Fig. 3c, the value of data element `date of birth` is degraded with a suitable offset following the execution of the activity *Check Age* and later degraded one additional level with a suitable offset after the execution of the activity *Assign to Event*.

Implementation of Intra-Process Data Degradation

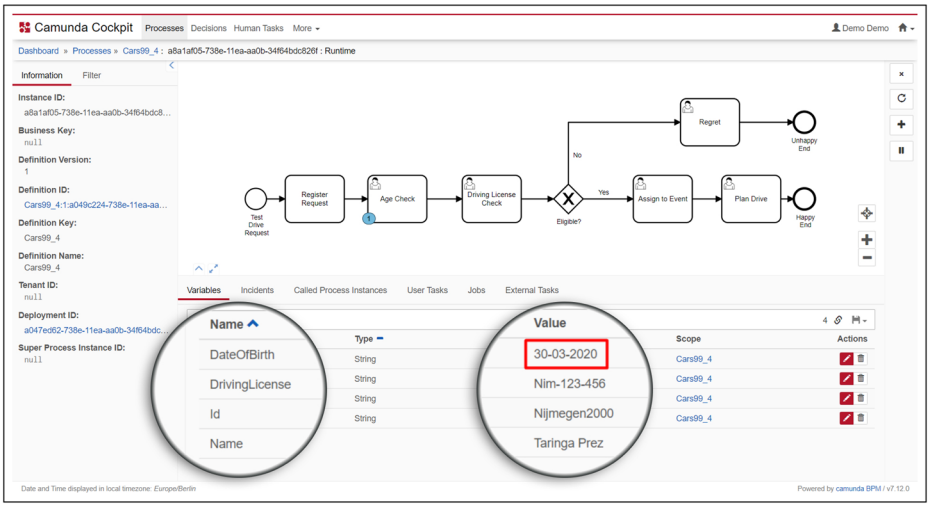
We have realised our intra-process data degradation approach in Camunda BPM[2], which is a java-based open-source workflow and decision automation platform. Camunda BPM currently supports BPMN (Business Process Modeling Notation), CMMN (Case Management Model and Notation) and DMN (Decision Model and Notation). Being a powerful workflow platform, the process engine can be bootstrapped in a java application or alternatively java applications can be deployed to the process engine through webapps mechanism. The engine provides Java APIs for providing interfaces.

In our implementation, the car dealership test drive business process (cf. Fig. 1) is coupled with the event-driven data degradation policy of Fig. 3b. The developed webapps are available on SurfDrive[3]. We are presenting screenshots of the Camunda *cockpit* at different process stages as Fig. (4a–c). The date of birth, represented as variable `DateOfBirth` in the screenshots, is provided at registration stage (cf. Fig. 4a). Once utilised for checking age eligibility for test drive by the activity *Age Check*, the `date of birth` data element is degraded one level such that only month and year of birth are retained (cf. Fig. 4b). If eligibility is proven in every respect, the request is further processed through the activity *Assign to Event* and `date of birth` is further degraded to contain only the birth year (cf. Fig. 4c).

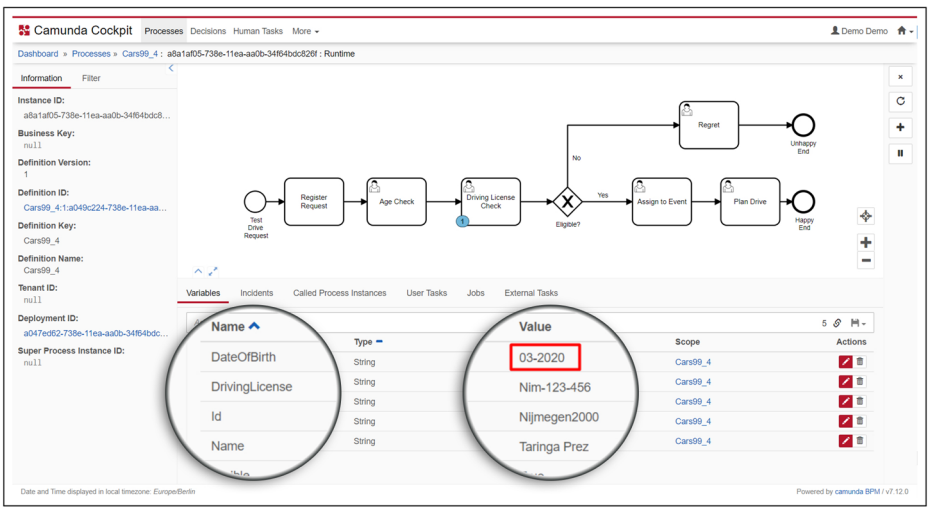
To illustrate the merits and effectiveness of our approach, we perform a comparative analysis of our approach with a conventional implementation of the business process of Fig. 1. We refer by "non-degradation approach" in the rest of this section to the approach that does not incorporate data degradation . We use the impact of a data breach in terms of the potential of the exposed data to reveal identification of the data subjects as a comparison metric. In the

[2] https://www.camunda.com

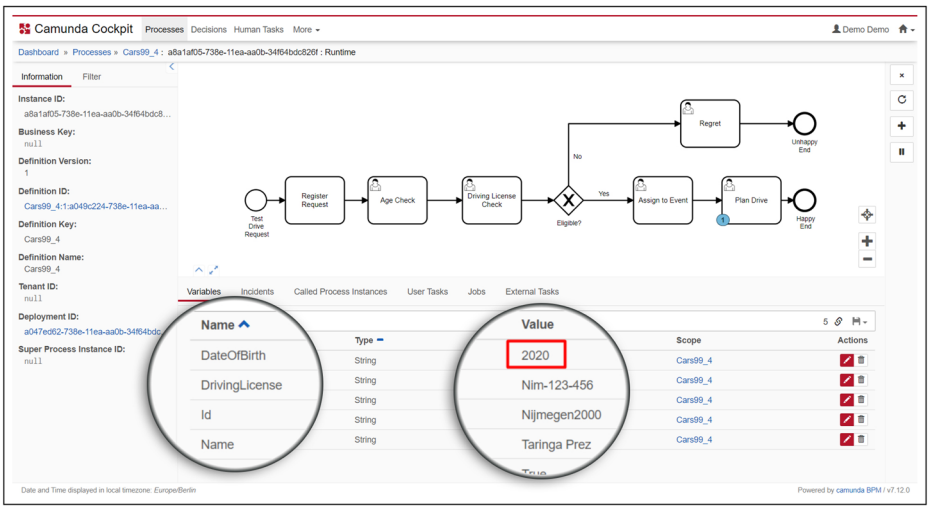
[3] https://surfdrive.surf.nl/files/index.php/s/E7mU4UQCLffmyoQ



(a) Process Variables Without Degradation.



(b) *Level1* Degradation of Data Element `date of birth` (cf. Figure 3b).



(c) *Level2* Degradation of Data Element `date of birth` (cf. Figure 3b).

Fig. 4 Screenshots of camunda process execution with degraded data element

non-degradation approach, the precision of process data is not reduced during the course of the process, and therefore data breach impact remains uniformly high throughout the process timeline.

For our approach, we consider data breaches happening at different instances in our process timeline. The impact of a data breach happening at any instance before the activity *Age Check* gets executed will be the same as the non-degradation approach, as the data is fully-precised yet. Next imagine a data breach happening at any instance in-between the execution of the activities *Age Check* and *Driving License Check*. The data breach impact of our approach will be relatively less as compared to the non-degradation approach since, as per our approach, the data element `date of birth` does not include the day of the birth. Further, a data breach happening at any instance after the execution of the activity *Assign to Event* in our approach will bear significantly less data breach impact in comparison to the non-degradation approach as the data has significantly lost its distinctive value due to the dropping of day and month of the birth details. The mentioned data breach impacts are also visualised in the coloring of the data degradation policies in Fig. 3.

From the process specification, we know that a complete instance of the business process of Fig. 1 may span over several months. Also, the *elapsed time* till the activity *Assign to Event* usually counts for a fraction of the whole process completion time. We can conclude that in the context of the processes bearing characteristics similar to our example business processes, our proposed intra-process data degradation approach makes the processing environment less vulnerable during the process life in comparison to the non-degradation approach.

Post-execution GDPR Compliance

This section details on the post-execution GDPR compliance with the business process traces from two distinct angles: ascertaining compliance to GDPR's data subject rights, and practicing data security in processing of event data for mining purposes. In the following, we detail the former in Sect. 4.1 and the latter in Sect. 4.1.1.

Enabling Compliance to Data Subject Rights

In this section we first detail on the GDPR's data subject rights. Then, we explain basic process mining tasks and their potential in ascertaining compliance with the GDPR's data subject rights in business processes. Finally, we provide an evaluation of our process mining based Right to be Forgotten (RTBF) compliance checking tool.

Relevant GDPR Articles

GDPR caters for the distrust of data subjects on data controllers regarding the possession and processing of their personal data. It empowers data subjects with a broad range of rights over their personal data. In Articles 6 and 7 declaring data subject's *consent* as a mandatory requirement for data processing, Articles 15 to 22 detail on the various data subject rights. Article 12 on one hand obligates data controllers to facilitate the exercise of data subject rights listed under the Articles 15 to 22, and on the other hand binds them to provide information on action taken on any request made under Articles 15 to 22 to the concerned data subject without undue delay.

Data subject rights are essentially data controller's obligations. In order to be able to fulfill data subject rights, data controllers might need to rework the control perspective of business processes either by changing business processes or by including additional behaviour into the process. For *without undue delay* aspect of fulfillment of data subject rights, the time perspective of business processes needs to be addressed. We will briefly explain these different business process perspectives in the next section.

Compliance Tools and Techniques

Process Mining Process mining is a discipline which is positioned between data mining and business process management and provides insights on the dynamics of the process execution. There are three main categories of process mining techniques: process model discovery, process conformance or compliance checking, and process enhancement. The two first-class citizens of process mining are *event logs* and *process models*, also known as observed behaviour and desired behaviour respectively.

An event log "L" consists of the process execution data recorded in the information systems. A single process instance, known as a *trace*, contains the (orderly) recorded sequence of events for each case (process instance). Each event shall at least contain the information tuple (case id, activity name, timestamp) for the primary process mining purposes. Additional event information results in discovery of interesting insights on the process.

A process model "$\mathcal{M}$" can be represented as a Petri net, $N=(P,T,F,A)$ where P in the tuple represents a finite set of Places, T is the set of Transitions, F is the set of arcs connecting places to transitions and transitions to places, and A is the set of Activities. Referring to Fig. 5, the middle row is a process model of lead management process of a car-dealership. Places in the model represent the different states of a process while transitions represent actions which result in changing the state of the process.

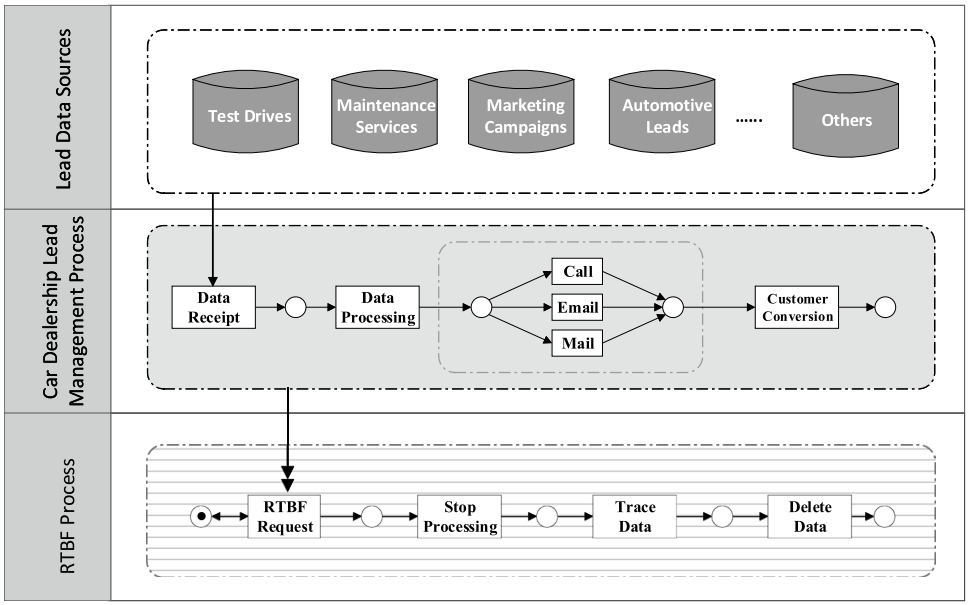


Fig. 5 RTBF-compliant car dealership lead management process model

Process discovery techniques take as input an event log "L" and discover a process model "$\mathcal{M}$". Business processes gets adapted under the influence of contextual factors and drifts. An event log data "L" contains the real execution data, therefore the discovered process model "$\mathcal{M}$" represents the de facto process model.

Conformance checking techniques gauge the harmony between process execution in real environment, i.e. as-is behaviour and the behaviour documented or perceived by the process owners, i.e. to-be behaviour. Conformance checking takes as input a process model "$\mathcal{M}$", be it discovered or provided by the process owners, and an event log "L". The two input entities are confronted to discover disagreements between traces in the event log and behaviour allowed as per the process model. Literature [1] details on many factors leading to deviations in execution of business processes.

Compliance checking techniques, a variant of conformance checking, check the compliance of the executed process instances (cases) against a behavioural constraint or requirement. The behavioural constraint is transformed into a formal representation with embedding the permissible behaviour. Process traces are evaluated against the behavioural constraint and deviations in the traces are diagnosed, if any.

Process mining discipline recognises three different perspectives on business processes: control-flow, data, and time. Control-flow perspective looks onto the orientation or spatial aspect of process activities in a business process. Data perspective counts on the data elements accessed, written, and updated by activities in a business process. Time perspective looks into temporal and performance aspect of process activities. To relate these perspectives to the GDPR, data miminisation is concerned with the data perspective of business processes. Data subject rights might necessitate changes in the control-flow perspective of business processes. The fulfillment of data subject rights *without undue delay* aspect may require reworking the time perspective of business processes.

Model Adaptation The implementation and the inclusion of data subject rights in the business processes bear varying implications. For some of these rights, a straightforward addition of a subprocess to the business process model may be sufficient. While others may necessitate reworking the primary business process as well. We are considering Article 17 of the GDPR i.e., *right to erasure*, which in our opinion is the most rigorous as it may even interrupt the execution of the primary processes. Titled as "Right to erasure" and commonly known as the "right to be forgotten" (RTBF), Article 17 requires that on receipt of an erasure request, the data controller should erase personal data of the concerned data subject without undue delay, except in few special cases.

In the context of the changes required for enabling RTBF compliance, refer bottom row of Fig. 5, RTBF process is structurally annexed to the business process of the car-dealership. From behavioural point of view, the RTBF process acts as a reset net [28] which takes control out of the business process upon triggering. This essentially stops the further processing of the data to be deleted. All the data repositories containing the data of the respective data subject are sorted out and all the concerned records therein are deleted. The actions taken by the data controller in response to the RTBF request must be acknowledged to the data subject as per Article 12. With the event data logged for the modified business process, our GDPR compliance checking techniques can ascertain RTBF compliance in the post-execution scenario, explained in coming section.

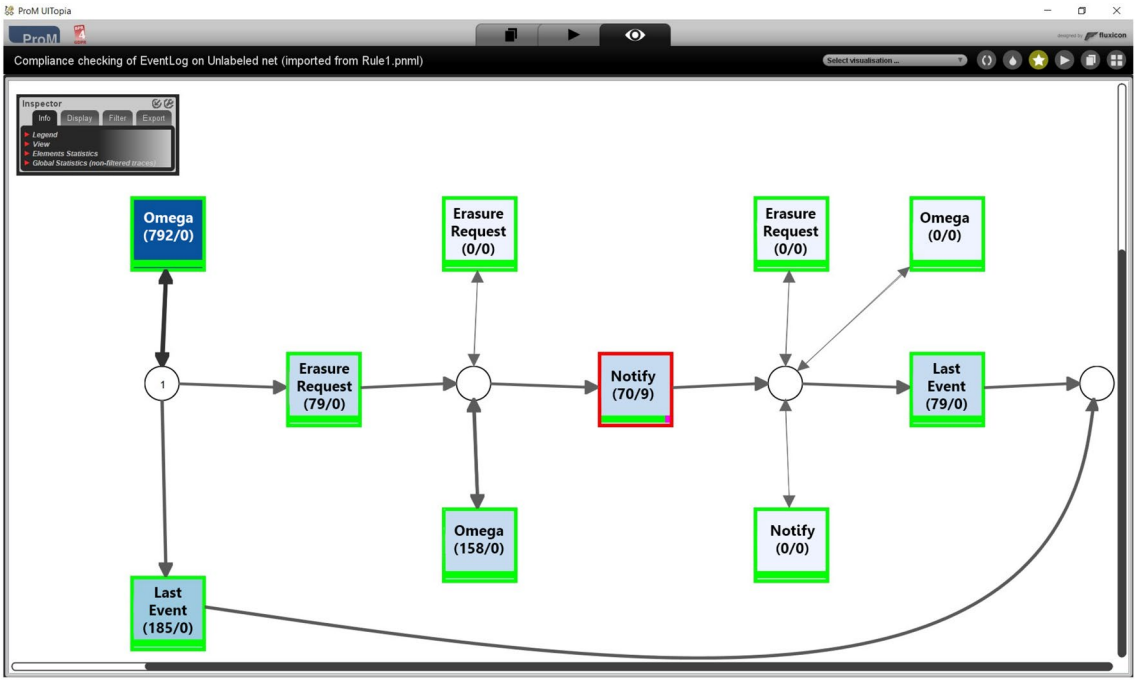


Fig. 6 Car dealership process GDPR compliance diagnosis

Implementation of RTBF-Compliance Process Mining Tool

We are considering the car dealership lead management process, the middle row of Fig. 5. Leads data from multiple sources, such as data from the test drive process of Fig. 1, is acquired by the lead management process. The acquired data is processed for identification of quality leads. The identified quality leads are contacted through multiple communication channels for customer conversion, i.e. selling, cross-selling or up-selling.

As discussed in the preceding section, we adapted the non-RTBF compliant process to RTBF compliant version. We simulated the RTBF compliant business process model of Fig. 5 through CPN Tools[4] for creating synthetic event log. The simulation is configured to embody controllable noise in the form of non-compliant traces. Our implementation of the RTBF-compliance checking technique within ProM [1] together with a synthetic event log and a step-by-step installation guide is available online[5].

The output of our RTBF compliance checking technique is provided through a Petri net representation as in Fig. 6. The *Erasure Request* transition represents the RTBF request by the data subject. The *Notify* transition represents the acknowledgment of the data deletion to the data subject as per Article 12. The red coloured perimeter of the transition *Notify* signals that deviations have been diagnosed at this activity. The statistics inside the transition provides the information that in 9 out of 79 RTBF cases the data subject was not communicated about action(s) taken as a result of his RTBF request. By clicking on this "problematic" transition, information about non-compliant traces can be filtered for further analysis.

In operational environments, events are recorded in massive volume. Analysis tools and techniques should therefore be scalable to deal with large volume of event data. We evaluated our RTBF compliance checking technique on event log of up to 500,000 cases. Evaluation was performed on a laptop with an Intel Core i7-7700HQ CPU, 32 GB of RAM and running Windows 10 operating system. Figure 7 presents the scalability of the compliance checking tool w.r.t. the increase of total number of cases considered in the log file. Our tool exhibit quasilinear time complexity in the number of considered cases.

Enabling Compliance to Data Security

In this section we highlight the GDPR articles related to the requirement of practicing data security in processing of personal data. We bridge the data security requirement with processing of event data in process mining techniques. Finally, we present our vision on utilising differential privacy technique in order to fulfill data security requirement in process discovery techniques.

[4] http://cpntools.org

[5] https://surfdrive.surf.nl/files/index.php/s/aUVFrw8GRccICpy

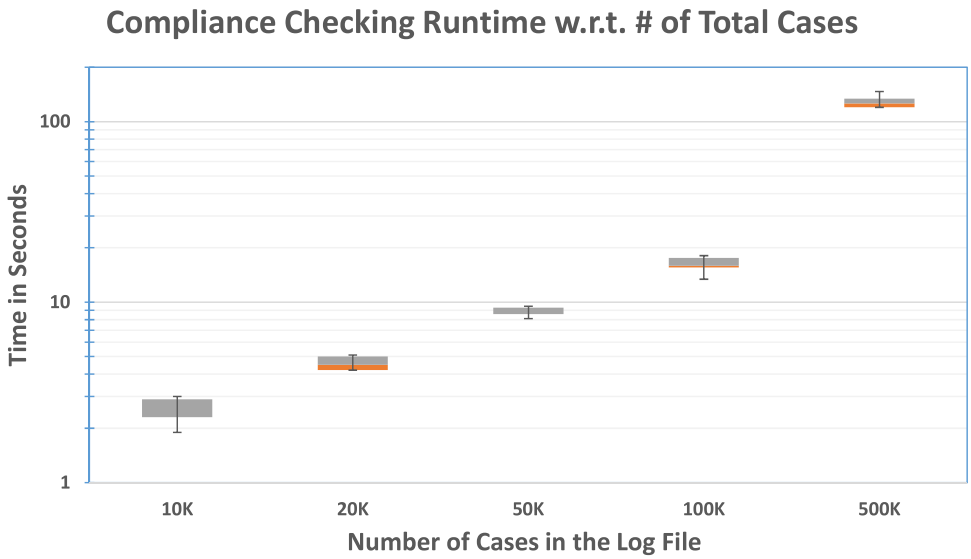


Fig. 7 The scalability of the RTBF compliance tool w.r.t. the number of cases in different log files of the *Lead Management* process

Relevant GDPR Articles

GDPR reiterates that personal data should be processed in a manner that ensures appropriate security and confidentiality of the data. While not precluding any other measures, pseudonymisation has been exemplified in GDPR at multiple occasions as an effective security technique. The 'integrity and confidentiality' processing principle of Article 5 requires that personal data should be processed with ensuring appropriate data security. Article 25 requires pseudonymisation to be part of the measures considered both at the time of the determination of the means for processing and at the time of the processing itself, i.e. privacy by design and default. Article 32 of the GDPR requires data controllers and processors to implement appropriate technical and organisational measures to ensure a level of appropriate data security by mentioning pseudonymisation and encryption as an example. Article 89 includes pseudonymisation into the safeguards and derogations related to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

Compliance Tools and Techniques

Process mining, in essence, is a process on its own that utilises process event data for almost all of its techniques. The event data may contain personal data of the data subjects as well as organisational resources related with the business process. Therefore, the data security articles mentioned in the previous section shall also be taken care of during the mining of business processes. In this work, we are focusing on process model discovery. We aim to devise a setup where process models can be discovered without the exposure of the precise information contained in the event data.

Differential privacy is an information sharing mechanism where aggregated information about groups of subjects in a dataset is shared, but with concealing information about the individual subjects. A differential privacy engine is entrusted to have a full access to the underlying data and this engine can be queried for different statistics like count, mean or variance over a group of data-related subjects. The engine in response returns the queried information with addition of some systematic noise such that the exact information, especially regarding individual records, cannot be revealed. The number of allowed queries is subjected to a certain privacy budget which gets depleted with the issuance of queries. A limitation of the number of queries through *privacy budget* ensures that it is not possible to reveal "precise" information about individual subjects by performing "high" number of queries.

Algorithmically, process discovery is usually a two-step process. At a first step, the event data is transformed into some intermediate representation which is utilised to construct a process model in a second step [37]. Directly-follows relations are an example of such intermediate representations. The quality of intermediate representation, in terms of true representation of the underlying event log, has a direct impact on the quality of the discovered process model. As per our approach, we aim to leverage differential privacy to build an acceptable intermediate representation in a privacy-abiding manner.

Referring to Fig. 8, we are specifying two environments in a process discovery setup. The actual process discovery algorithm and its querying agent are considered to be residing in a non-trusted environment while the event log and the privacy engine reside in the trusted environment. The privacy engine should have access to the event log for extracting statistical information of the events. The process discovery querying agent, residing in the non-trusted environment,

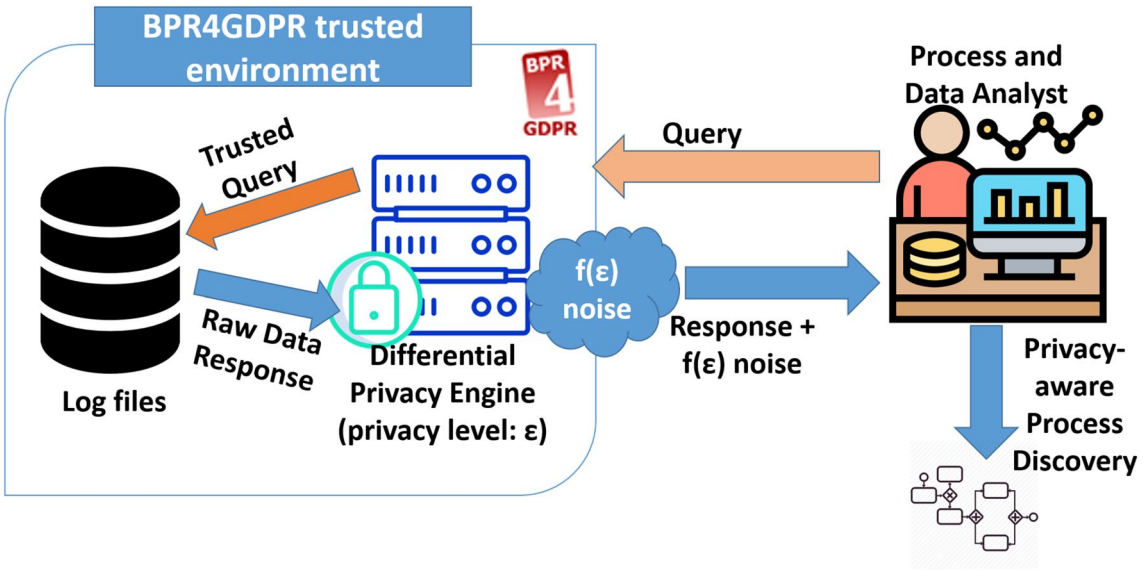


Fig. 8 Our proposed setup for applying differential privacy in process discovery

can only communicate with the differential privacy engine in the trusted environment by issuing queries in order to acquire required information on the event data.

The semantics of the query depends on the intermediate representation utilised by the process discovery. For instance, in case of directly-follows relation, the process discovery querying agent will query the number of instances where an activity A is directly-followed by an activity B. Mainly, the *count* feature of differential privacy will be utilised. The privacy engine in response to the query provides the relevant statistics along with addition of ϵ noise. As with any differential privacy setup, the process discovery querying agent can issue up to a certain number of queries to the differential privacy engine under the constraint of assigned privacy budget.

Existing techniques like [25], suffer from in-efficient querying mechanism and discovery of low-quality process models. We aim to effectively utilise the querying budget to discover high quality process models in the real time [15].

Challenges and Future Work

In this work, we presented solutions for enabling GDPR compliance with respect to data minimisation, data subject rights in business processes, and data security while mining processes. In this section we present the major challenges and future directions regarding these three areas.

Our intra-process data degradation implementation proved the initial effectiveness of the introduced concepts. Nevertheless, some challenges are foreseen. The complexity of the processing environments is considered to be challenging for our intra-process data degradation approach. Most often, same data are processed in multiple processes concurrently, sequentially or in a cascading pattern. As another scenario of processing environment's complexity, multiple organisations may be involved in beyond-boundary processing as part of a single process. The structural complexity of real world business processes is another challenge for intra-process data degradation. Process models contain parallelism, choices and the most challenging looping constructs. All these constructs introduce behavioural complexity in the incumbent business process. Our current versions of the data degradation policies are still unable to handle such behavioural diversity and are therefore making data degradation hard to realise in such processes. As a future work in the direction of data minimisation, we will investigate much more advanced, probably hybrid, data degradation policies.

We presented a solution for enabling GDPR compliance of business processes with respect to the data subject rights, complemented by process mining discipline. Particularly, we elaborated on ascertaining RTBF-compliance using a novel tool. The complexity of the processing environments and the structural complexity of real world business processes is equally challenging in this direction. Apart from detecting deviations, an automated feedback in the form of remedial actions should also be provided to close the loop.

For the realisation of data security in mining business processes, we also see a lot of challenges. Effective query-synthesis is one of the major challenge in differentially-private process mining. Some process mining algorithms require information on the level of activities, such as Directly-Follows Graphs (DFGs), while others require

information on the level of process traces for the discovery of meaningful process models. A frugal querying mechanism needs to be devised such that the information requested from the differential privacy agent is on one hand minimal and on the other hand sufficient for all of these process discovery algorithms. Despite the fact that the discovery process in such a setup is based on noisy statistics, maximising the utility of the discovered process models is a further challenge to be addressed.

Funding The authors have received funding within the BPR4GDPR project from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 787149.

References

1. van der Aalst WMP. Process mining: data science in action. 2nd ed. Berlin: Springer; 2016. https://doi.org/10.1007/978-3-662-49851-4.
2. Antignac T, Le Métayer D. Trust driven strategies for privacy by design. In: IFIP international conference on trust management, IFIPTM 2015. Cham: Springer; 2015, pp. 60–75.
3. Anciaux N, Nguyen B, Vazirgiannis M. Miminum exposure in classification scenarios; 2011.
4. Anciaux N, Boutara D, Nguyen B, Vazirgiannis M. Limiting data exposure in multi-label classification processes. Fundamenta Informaticae. 2015;137(2):219–36.
5. Anciaux N, Nguyen B, Vazirgiannis M. Limiting data collection in application forms: a real-case application of a founding privacy principle. In: 2012 tenth annual international conference on privacy, security and trust, PST 2012. IEEE; 2012, pp. 59–66.
6. Hilderman RJ, Hamilton HJ, Cercone N. Data mining in large databases using domain generalization graphs. J Intell Inf Syst. 1999;13(3):195–234.
7. Anciaux N, Bouganim L, Van Heerde H, Pucheral P, Apers PMG. Data degradation: making private data less sensitive over time. In: Proceedings of the 17th ACM conference on information and knowledge management, CIKM 2008; 2008, pp. 1401–2.
8. Anciaux N, Bouganim L, Van Heerde H, Pucheral P, Apers PMG. Instantdb: enforcing timely degradation of sensitive data. In: 2008 IEEE 24th international conference on data engineering, ICDE 2008. IEEE; 2008, pp 1373–5.
9. Anciaux N, Bouganim L, Van Heerde H, Pucheral P, Apers P. The life-cycle policy model; 2008.
10. van Heerde HJW, Anciaux NLG, Fokkinga MM, Apers PMG. Exploring personalized life cycle policies. CTIT Technical Report Series Supplement/TR-CTIT-07-85; 2007.
11. van Heerde HJW, Anciaux N. Data degradation to enhance privacy for the Ambient Intelligence. CTIT Technical Report Series 11/06-74; 2006.
12. Geambasu R, Kohno T, Levy AA, Levy HM. Vanish: increasing data privacy with self-destructing data. In: USENIX security symposium, vol. 316; 2009.
13. Pearson S, Casassa-Mont M. Sticky policies: an approach for managing privacy across multiple parties. Computer. 2011;44(9):60–8.
14. Leemans SJJ, Fahland D, Van Der Aalst WMP. Process and deviation exploration with inductive visual miner. In: International conference on business process management, BPM 2014 (Demos) 1295, no. 46; 2014.
15. Hassani M, Siccha S, Richter F, Seidl T. Efficient process discovery from event streams using sequential pattern mining. In: 2015 IEEE symposium series on computational intelligence, SSCI 2015. IEEE; 2015, pp. 1366–73.
16. Hassani M. Concept drift detection of event streams using an adaptive window. In: 33rd international ECMS conference on modelling and simulation, ECMS 2019; 2019, pp. 230–9.
17. Adriansyah A, van Dongen BF, van der Aalst WMP. Conformance checking using cost-based fitness analysis. In: 2011 IEEE 15th international enterprise distributed object computing conference, EDOC 2011. IEEE; 2011, pp. 55–64.
18. Carmona J, van Dongen B, Solti A, Weidlich M. Conformance checking: relating processes and models. Springer; 2018.
19. Ramezani E, Fahland D, van der Aalst WMP. Where did I misbehave? Diagnostic information in compliance checking. In: International conference on business process management, BPM 2012. Berlin, Heidelberg: Springer; 2012, pp. 262–78.
20. Ramezani E, Fahland D, van der Aalst WMP. Supporting domain experts to select and configure precise compliance rules. In: International conference on business process management, BPM 2013, pp. 498–512. Cham: Springer; 2013.
21. Taghiabadi ER, Gromov V, Fahland D, van der Aalst WMP. Compliance checking of data-aware and resource-aware compliance requirements. In: OTM confederated international conferences on the move to meaningful internet systems, OTM 2014. Berlin, Heidelberg: Springer; 2014, pp. 237–57.
22. Fahrenkrog-Petersen SA, van der Aa H, Weidlich M. Pretsa: event log sanitization for privacy-aware process discovery. In: 2019 international conference on process mining, ICPM 2019. IEEE; 2019, pp. 1–8.
23. Ciriani V, De Capitani Di Vimercati S, Foresti S, Samarati P. K-anonymity. In: Secure data management in decentralized systems. Boston, MA: Springer; 2007, pp. 323–53.
24. Sweeney L. k-anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl-Based Syst. 2002;10(05):557–70.
25. Mannhardt F, KoschmiderA Baracaldo N, Weidlich M, Michael J. Privacy-preserving process mining. Bus Inf Syst Eng. 2019;61(5):595–614.
26. Dimitrova D, De Hert P. The right of access under the police directive: small steps forward. In: Annual privacy forum, APF 2018. Cham: Springer; 2018, pp. 111–30.
27. Agarwal S, Steyskal S, Antunovic F, Kirrane S. Legislative compliance assessment: framework, model and GDPR instantiation. In: Annual privacy forum, APF 2018. Cham: Springer; 2018, pp. 131–49.
28. Dufourd C, Finkel A, Schnoebelen P. Reset nets between decidability and undecidability. In: International colloquium on automata, languages, and programming, ICALP 1998, pp. 103–15. Berlin, Heidelberg: Springer; 1998.
29. Zaman R, Cuzzocrea A, Hassani M. An innovative online process mining framework for supporting incremental gdpr compliance of business processes. In: 2019 IEEE international conference on big data, IEEE Big Data 2019. IEEE; 2019, pp. 2982–91.

30. Zaman R, Hassani M. Process mining meets GDPR compliance: the right to be forgotten as a use case. In: 2019 international conference on process mining doctoral consortium, ICPM-DC 2019. CEUR-WS.org; 2019.
31. Lioudakis GV, Koukovini MN, Papagiannakopoulou EI, Dellas N, Kalaboukas K, de Carvalho RM, Hassani M, et al. Facilitating GDPR compliance: the H2020 BPR4GDPR approach. In: Conference on e-Business, e-Services and e-Society, I3E 2019. Cham: Springer; 2019, pp. 72–8.
32. McSherry FD. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD international conference on management of data, SIGMOD 2009; 2009, pp. 19–30.
33. Dwork C. Differential privacy: a survey of results. In: International conference on theory and applications of models of computation, TAMC 2008. Berlin, Heidelberg: Springer; 2008, pp. 1–19.
34. Wilson RJ, Zhang CY, Lam W, Desfontaines D, Simmons-Marengo D, Gipson B. Differentially private sql with bounded user contribution. In: Proceedings on privacy enhancing technologies 2020, PETS 2020, no. 2; 2020, pp. 230–50.
35. Holohan N, Braghin S, Aonghusa PM, Levacher K. Diffprivlib: the IBM differential privacy library; 2019. arXiv:1907.02444.
36. Baskar K, Hassani M. Online comparison of streaming process discovery algorithms. In: 2019 dissertation award, doctoral consortium, and demonstration track at BPM, BPMT 2019. CEUR-WS.org; 2019, pp. 164–8.
37. Hassani M, van Zelst SJ, van der Aalst WMP. On the application of sequential pattern mining primitives to process discovery: overview, outlook and opportunity identification. Wiley interdisciplinary reviews: data mining and knowledge discovery 9, no. 6, e1315; 2019.