Real-Time Anonymization of Sensitive Personal Data Using a Service-Based Architecture

Fabio Giampaolo[1], Stefano Izzo[1], Stefano Siccardi[2], Antongiacomo Polimeno[3], Valerio Bellandi[3] and Francesco Piccialli[1*]
[1] Department of Mathematics and Applications ”R. Caccioppoli”, University of Naples Federico II, Italy
{fabio.giampaolo,stefano.izzo,francesco.piccialli}@unina.it
[2] Consorzio Interuniversitario Nazionale per l’Informatica, Italy
stefano.siccardi@consorzio-cini.it
[3] Department of Computer Science, Università degli Studi di Milano, Italy
valerio.bellandi@unimi.it

***Abstract*—Anonymization is an important aspect of data privacy protection, especially in the context of sensitive personal information collected through sensors. In this paper, we propose a new service-based architecture for anonymizing such data in real-time, ensuring that data is accessible to authorized users while maintaining privacy. Our architecture is based on the annotation of data at ingestion time, where privacy levels are assigned to sets of columns. The anonymization procedure is performed by compressing and encoding the data through an autoencoder model, where the encoder and decoder functions are defined as parametric functions composed of multiple hidden layers.**

***Index Terms*—Data Infrastrucure, Anonymization, Healthcare Dataset, Privacy, Autoencoder.**

I. Introduction

In recent years, there has been a significant increase in data generation due to advances in technology and widespread connectivity. This surge in data collection encompasses various aspects of our lives, including social media platforms and e-commerce websites. Consequently, businesses and governments have begun harnessing this vast amount of data for analytical purposes, enabling them to gain valuable insights and make informed decisions [4]. However, the rapid growth in data generation has also raised concerns regarding data privacy and security. In response, governments worldwide have implemented regulations like the General Data Protection Regulation (GDPR) to safeguard individuals’ data privacy. In the last years the *service architecture* emerged as a methodology to manage this data acquired by devices. This has provided a framework for managing and securing data acquired from devices. This architectural pattern involves developing software applications composed of individual services that can be combined to form a complete application. Each service offers specific functionality accessible through a well-defined interface. In healthcare data management, for example, a service architecture can offer a modular approach to data management and analysis. Among the various techniques employed for privacy preservation, data anonymization is one commonly used method. The literature proposes numerous works on anonymization techniques in data management systems, often based on principles such as k-anonymity, l-diversity, and t-closeness. Data anonymization plays a critical role in maintaining privacy in the era of big data. While the k-anonymity, l-diversity, and t-closeness models have gained widespread adoption, they come with high computational complexity and are susceptible to various re-identification attacks.

A. Generalities

We list some control requirements, that should be met by any system dealing with sensible data, both related to healthcare and to other subjects.

- R1: Right to access the data must be evaluated before data analytics takes place.
- R2: As federations within big data ecosystem must be considered, authentication should be managed by a separate and integrated module [1]
- R3: data must be properly protected and shared only to authorized users and for authorized operations. Access control must protect data during their entire life cycle
- R4: Fine-grained access control must be supported, dealing with both structured and unstructured data. In particular, when structured data are considered, policies can refer to a single cell, a column, a tuple or an entire table of structured data.
- R5: Access control enforcement should not use data ownership as the only attribute to define access rights. It should be applied at ingestion time on the basis of a flexible set of characteristics of the specific data context.
- R6: As a consequence, access control should be driven by dynamic and contextual annotations on data
- R7: : Access control should be highly efficient and scalable to cope with the increasing cardinality of data and rate of requests.

Privacy issues should be considered at design time, and access control should be embedded at all levels of the architecture (see e.g. [2]). In this work we focus our attention on the Data Acquisition, Storage and Sharing and Privacy Services with the objective to propose an architecture that supports the release of data anonymously based on the level of the user’s authorization. Fig. 1 describes the main pipeline of the system.

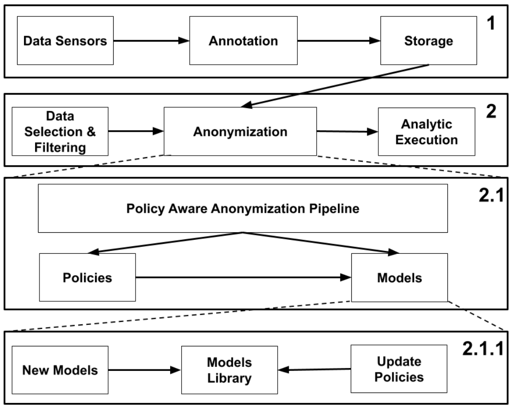

Fig. 1. Data Elaboration Pipeline

B. Anonymization as a Service

The proposed anonymization model can be integrated into a service context to provide various anonymization levels based on a set of policies. A module for managing the creation and storage of models trained on different datasets and for different anonymization features (as defined by some policies) can be designed as outlined in this section.

Given the generic dataset k, a meta-model $\mathcal{M}_k$ storing information about the number and the type of features is generated. The policies $\mathcal{P}_k$ related to dataset k specify different required levels of anonymization, i.e. groups of features that need to be anonymous based on access levels to the dataset. For each of the n required anonymization levels, n models are generated for dataset k, each one anonymizing a different group of features. Mathematically, let $\mathcal{D}_k$ represent the dataset and F_k its features. Sub-groups of features $F_k^i \subset F_k$ are selected, and models $\mathcal{M}_k^i$ are trained to produce the anonymous version of the dataset $\mathcal{D}_k^i = \mathcal{M}_k^i(\mathcal{D}_k|F_k^i) \cup (\mathcal{D}_k \setminus \mathcal{D}_k|F_k^i)$.

These models can be stored and called upon whenever a request based on the predetermined policies is made, allowing each request to be processed in *near real-time*. Expiration conditions can also be assigned to these models. If an incoming request does not meet the policies $\mathcal{P}_k$, meaning that anonymization is needed for a group of features not covered by the standard policies for the k-th dataset, a temporary policy can be generated. The anonymization models so created *on demand* can be discarded or saved.

Architectural Overview

In order to support the anonymization as a service, an architecture has been implemented following the aforementioned guidelines. The *access control* and *user management* modules implement the usual login functions and general permissions to see and manage the data. With *client components* we mean generic processes that need to access the data, like front end interfaces for data exploration and visualization or Machine Learning / Analytics components. In the ingestion phase, we highlight the *data annotation* step, that links raw data to features and therefore to proper anonymization models.

The data anonymization model

The procedure of data anonymization is achieved by compressing and encoding the features of interest of the dataset through the encoder of an AutoEncoder (AE) model, a Neural Network architecture defined by two submodules, an encoder and a decoder, defined as one or a composition of multiple parametric functions and in this case constructed using Dense layers.

$$\begin{aligned} \nu(\cdot; W, b) : \mathbb{R}^o &\longrightarrow \mathbb{R}^p \\ v &\longmapsto w = \sigma(Wv + b) \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{p \times o}$, $b \in \mathbb{R}^p$, o and p dimensions of the input and the output of the layer ν.The term $Wv + b$ is a linear aggregation function, while σ is named *activation function* and it is generically a non-linear function.

With this definition we can denote both the encoder and the decoder functions as the functions g and h, respectively, defined as follows:

$$\begin{aligned} g : \mathbb{R}^m &\longrightarrow \mathbb{R}^n \\ x &\longmapsto z = g_l \circ g_{l-1} \circ \cdots \circ g_2 \circ g_1(x), \end{aligned} \tag{2}$$

$$\begin{aligned} h : \mathbb{R}^n &\longrightarrow \mathbb{R}^m \\ z &\longmapsto \hat{x} = h_l \circ h_{l-1} \circ \cdots \circ h_2 \circ h_1(z). \end{aligned} \tag{3}$$

The encoder g, composed by $g_1, \ldots, g_l$ hidden layers, takes in input a row of the dataset x with number of features of interest m and returns a smaller dimensional data z with dimension n. The decoder h takes in input z and returns a vector $\hat{x}$, s.t. $dim(x) = dim(\hat{x})$. The codomain of the encoder function is called *latent space*. The AE model can be so defined as:

$$\begin{aligned} f : \mathbb{R}^m &\longrightarrow \mathbb{R}^m \\ f : x &\longmapsto \hat{x} = h(z) = (h \circ g)(x). \end{aligned} \tag{4}$$

whose aim is to extract a smaller dimensional representation of the data while also making it possible to closely reconstruct the dataset, thereby preserving the original properties that could be are exploited by statistical analysis.

II. Experiment and Evaluation

Our proposal has been tested healthcare-related datasets to confirm its applicability in such a scenario, and also on datasets containing general demographic features to assess its generality. The data anonymization system has been tested by assessing the performances of a ML model in terms of classification tasks through the metrics Accuracy, Precision (micro if multi-class), Recall (micro if multiclass) and F1-score, before and after the anonymization. In particular, we applied such a procedure on the data sets *Diabetes*, *Obesity levels*, *Adult* and *Credit Card* , all provided by the UCI Machine Learning Repository [5].

Table I shows promising results, with the highest performance metric reduction in percentage being 5% circa, indicating that the encoding procedure is capable of extrapolating the distinctive patterns necessary for the ML models. As regards the general contexts, similar tests have been conducted on the

TABLE I

CLASSIFICATION RESULTS OF THE ML MODEL BEFORE AND AFTER THE ANONYMIZATION

	Diabetes		Obesity		Adult		Credit card	
	Original	Anonymized	Original	Anonymized	Original	Anonymized	Original	Anonymized
Accuracy	0.56	0.57	0.96	0.90	0.85	0.83	0.83	0.82
Precision	0.48	0.53	0.94	0.88	0.79	0.77	0.76	0.75
Recall	0.41	0.39	0.95	0.89	0.78	0.74	0.66	0.64
F1-score	0.40	0.36	0.96	0.88	0.79	0.75	0.68	0.67

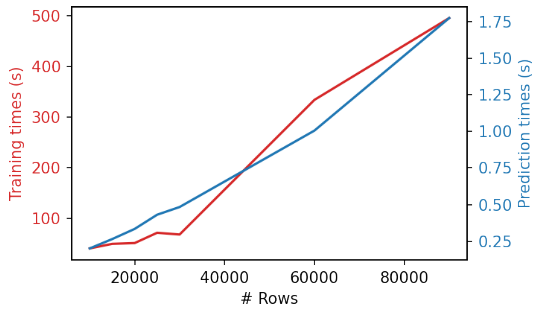

Fig. 2. Behaviour of the training and anonymization times as the number of the rows grows in the case of *Adult* dataset

TABLE II

TIME REQUIRED TO TRAIN A MODEL AND THEN OBTAIN ANONYMIZED DATA FROM IT AS THE NUMBER OF ROWS GROWS. THE NUMBER OF ORIGINAL FEATURES IN 14.

	Adult	
# Rows	Training Time (s)	Anonymization Time (s)
10000	40.41	0.20
15000	48.91	0.26
20000	50.74	0.34
25000	59.83	0.41
30000	69.48	0.50
60000	339.68	0.96
90000	502.81	1.43

dataset *Adult*. As can be noticed from Table II and Figure 2, the general behaviour of the system, as the number of rows of the dataset increases, remains the same, exhibiting a linear growth.

III. CONCLUSION AND FUTURE WORK

In conclusion, service infrastructure is an essential aspect of the modern era of technology, enabling individuals and organizations to access a vast range of services, applications, and data. However, the increasing reliance on service infrastructure has also raised significant privacy concerns, as users' personal data is collected, stored, and processed by service providers. In this work we propose an infrastructure that allows to manage sensitive data by anonymizing information on the fly based on the level of the user and the kind of data acquired.

ACKNOWLEDGMENT

This work has been supported by the projects: i) FAIR: - Future AI Research (PE00000013), Spoke 3, under the PNRR MUR program ii), MUSA – Multilayered Urban Sustainability Action, under the National Recovery and Resilience Plan (NRRP) iii) SERICS projects (PE00000014) under the MUR NRRP funded by the EU - NextGenerationEU.

REFERENCES

[1] Anisetti, M., Ardagna, C. A., Braghin, C., et al. (2021). Dynamic and Scalable Enforcement of Access Control Policies for Big Data. In *Proceedings of the 13th International Conference on Management of Digital EcoSystems (MEDES '21)* (pp. 71-78). ACM.

[2] Perera, C., McCormick, C., Bandara, A. K., et al. (2016). Privacy-by-Design Framework for Assessing Internet of Things Applications and Platforms. In *6th Int. Conf. on the Internet of Things (IoT '16)*.

[3] Bellandi, V., Ceravolo, P., Damiani, E., et al. (2022). Smart Healthcare, IoT and Machine Learning: A Complete Survey. In *Hand. of Artificial Intelligence in Healthcare*, Int. Sys. Ref. Lib., vol 212, 99-126.

[4] Azzini, A., Barbon Jr, S., Bellandi, V., et al. (2021). Advances in Data Management in the Big Data Era. In *IFIP Advances in Information and Communication Technology, 600*, 99-126.

[5] Dua, D., and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[6] Bellandi, V., Ceravolo, P., Damiani, E., et al. (2022). Smart Healthcare, IoT and Machine Learning: A Complete Survey. In *Hand. of Artificial Intelligence in Health: Vol 2: Practicalities and Prospects*, 307-330.

[7] Peretokin, V., Basdekis, I., Kouris, I., et al. (2022). Overview of the SMART-BEAR Technical Infrastructure. In *Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2022)*, 117-125.

[8] Sweeney, L. (2002). K-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570.

[9] Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J. (2016). Beyond k-Anonymity: l-Diversity and t-Closeness. In *Datab. Anonym.*, Springer.

[10] Vinogradov, S., and Pastsyak, A. (2012). Evaluation of Data Anonymization Tools. In *The 4th Int. Conference on Advances in Databases, Knowledge, and Data Applications DBKDA*.

[11] Chen, M., Cang, L.S., Chang, Z., et al. (2023). Data anonymization evaluation against re-identification attacks in edge storage. *Wir. Net.*.

[12] Di Cerbo, F., and Trabelsi, S. (2018). Towards Personal Data Identification and Anonymization Using Machine Learning Techniques. In *New Trends in Databases and Information Systems. ADBIS 2018*, Springer.