


Tailored IoT & BigData Sandboxes and Testbeds for Smart,
Autonomous and Personalized Services in the European
Finance and Insurance Services Ecosystem



D3.14 – Data Governance Frameworks and Tools III

Revision Number	0.1
Task Reference	T3.5
Lead Beneficiary	GRAD
Responsible	Inés Ortega Fernández
Partners	GRAD, ATOS, JSI
Deliverable Type	Demonstrator (Dem)
Dissemination Level	Public (PU)
Due Date	2022-03-31
Delivered Date	2022-03-31
Internal Reviewers	NUIG, AGRO
Quality Assurance	INNOV
Acceptance	WP Leader Accepted and/or Coordinator Accepted
EC Project Officer	Beatrice Plazzotta
Programme	HORIZON 2020 - ICT-11-2018
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement no 856632

Contributing Partners

Partner Acronym	Role ¹	Author(s) ²
GRAD	Lead Beneficiary	Inés Ortega Fernández Eva Sotos Martínez Iago Abad Fernández
ATOS	Contributor	Nuria Ituarte Aranda
JSI	Contributor	Maja Škrjanc Filip Koprivec
NUIG	Internal Reviewer	Martin Serrano
AGRO	Internal Reviewer	Gregory Mygdakos
INNOV	Quality Assurance	John Soldatos

Revision History

Version	Date	Partner(s)	Description
0.1	2021-12-15	GRAD	ToC Version
0.2	2022-02-01	GRAD	Added GRAD contribution to Section 2
0.3	2022-03-01	GRAD	Added GRAD contribution to Section 3
0.4	2022-03-08	ATOS	Added ATOS contribution to Sections 2 and 3
0.5	2022-03-08	GRAD	Added first version of Section 1 and Conclusions
0.6	2022-03-14	JSI	Added JSI contribution to Sections 2
0.7	2022-03-15	JSI	Added JSI contribution to Section 3 and Conclusions
1.0	2022-03-16	GRAD, ATOS, JSI	Version for Internal Review
2.0	2021-03-25	GRAD, NUIG, AGRO	Version for Quality Assurance
3.0	2022-03-30	GRAD, INNOV	Version for Submission

¹ Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

² Can be left void

Executive Summary

This document introduces the INFINITECH's Data Governance considerations, addresses regulatory frameworks, and explains how Data Governance Framework and Tools were developed through the INFINITECH project to ensure the appropriate implementation regarding the use of Big Data and Data Analytics frameworks. To this end, a revision about big data characteristics in relation to their impact in the digital finance is presented and reported following a shortlist of identified challenges:

- 1) The increment of the amount of data handled in Big Data contexts constitutes a challenge for managing data privacy;
- 2) The velocity, variety and complexity of the data make necessary the use of automated tools that help manage the risks and regulatory requirements associated with such contexts.
- 3) Management of user identities is also challenging in Digital Finance: FinTechs need to onboard users into their platforms in a secure and reliable manner, using mechanisms that replace the need of manually inspecting physical identity documents.
- 4) The General Data Protection Regulation (GDPR) [1], depending on the nature of the data different data privacy and security measures need to be implemented; and
- 5) Pseudonymization is considered a security measure to allow the processing of personal data, while anonymization makes the individuals within a dataset non-identifiable, and therefore anonymised data is not considered personal data anymore.

In the context of the INFINITECH project, Data Governance is an essential mechanism to guarantee data security and privacy, as well as to establish the required workflows to manage data and information in companies, and in particular in banks, FinTechs and other insurance and financial organisations.

To meet Data Governance goals, the INFINITECH project provides the following tools and mechanisms:

- a) **A pseudonymization tool:** This tool provides mechanisms to pseudonymize unique identifiers, and the generalisation of numeric and time-stamped enriched transactional data by exploiting different techniques;
- b) **A tool for anonymizing data:** It provides an automatic framework for anonymizing datasets by automatically selecting the best configuration which fits the aspired privacy and utility goals, and
- c) **A mobile digital user onboarding service:** The tool, namely the Digital User Onboarding Services (DUOS) - provides a mobile service that allows the creation of virtual identities by combining digital certificates from government issued electronic IDs or passports with face images.

This deliverable contains the final design specifications initially described in D3.12 "Data Governance Frameworks and Tools I" [2] and D3.13 "Data Governance Frameworks and Tools II" [3] to their last version. In it, it is also presented the final implementation decisions included in each tool since the last deliverable as well as an analysis of the degree of accomplishment of the objectives of the task, and the achieved Technology Readiness Level [4]. To finalise the document, a set of conclusions is included in order to summarise the most important concepts and the topics addressed.

Table of Contents

Introduction	8
Objective of the Deliverable	8
Updates to the previous version	8
Insights from other Tasks and Deliverables	9
Structure	10
Final design of the Data Governance Framework and tools	11
Data pseudonymization tool	11
Data anonymization	12
Operation modes	13
Command Line Interface	14
REST API mode	15
Digital User Onboarding Tool	17
Description of final implementation of data governance tools	19
Data pseudonymization service	19
REST API specification	19
Data anonymization	20
Anonymization Operations	20
Privacy and utility metrics	22
Data Sources	22
Static datasets	23
Data in streaming	23
Digital user onboarding services tool	24
Conclusions	28
Appendix A: Literature	29
Appendix B: Anonymization configuration	30
Appendix C: Pseudonymization configuration	33

List of Figures

Figure 1 - pseudonymization tool conceptual diagram	12
Figure 2 – example execution of the CLI Configuration Tool	14
Figure 3 – metric’s results for each configured anonymization operation	15
Figure 4 - DUOS enrolment use case	18
Figure 5 - DUOS authentication use case	18
Figure 6 - workflow of data anonymization in streaming	24
Figure 7 - DUOS app initial screen (left) and new identity creation menu (right)	25
Figure 8 - DUOS waiting for a MRZ zone (left) and DUOS asking for ID reading NFC (right)	25
Figure 9 - DUOS virtual identify created (left) and virtual identities available (right)	26
Figure 10 - DUOS face recognition starting (left) and face verification process completed(right)	26
Figure 11 - DUOS scanning a SP QR code (left) and QR scan process completed (right)	27

List of Tables

Table 1 - Availability of the Data Governance Frameworks and reference Pilot	8
Table 2 - Description of supported anonymization operations	21

Abbreviations/Acronyms

Abbreviation	Definition
4AMDL	Anti-Money Laundering Directive IV
AES	Advanced Encryption Standard
AI	Artificial Intelligence
API	Application Programming Interface
APK	Android Application Package
BIC	Bank Identifier Code
BOS	Bank of Slovenia
CAK	Constrained Attacker K
CICD	Continuous Integration / Continuous Deployment
CLI	Command Line Interface
CSV	Comma Separated Values
DNI	National Identity Document (Spain)
DPO	Data Protection Orchestrator
DUOS	Digital User Onboarding System
eMRTD	Electronic Machine Readable Travel Document
EU	European Union
GDPR	General Data Protection Regulation
GPS	Global Position System
HPC	High Performance Computing
HTTP	Hypertext Transfer Protocol
IAM	Identity and Access Management
IBAN	International Bank Account Number
ICAO	International Civil Aviation Organization
ID	Identification
ILM	Information Loss Measure
IoT	Internet of Things
JSON	JavaScript Object Notation
MAE	Mean Absolute Error
MiFiD	Markets in Financial Instruments Directive
ML	Machine learning
MRZ	Machine Readable Zone

MSE	Mean Squared Error
MV	Mean Variation
NFC	Near Field Communication
OCR	Optical Character Recognition
PALMS	Platform for Anti Money Laundering Supervision
PSD	Payment Services Directive
QR	Quick Response
RA	Reference Architecture
RCAK	Random Constrained Attacker K
REST	Representational State Transfer
SHARP	Smart, Holistic, Autonomous, Regulatory Compliance and Personalised
SP	Service Provider
SQL	Structured Query Language
TRL	Technology Readiness Level
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VAT	Value Added Tax
VID	virtual Identity
WP	Work Package

1 Introduction

This document is the last deliverable of a series of three, related to the Data Governance Frameworks and Tools. The main goal is to present the definitive version of the data governance mechanisms that have been developed within the context of the INFINITECH project during the 26 months of the task “T3.5 Data Governance Mechanisms”. In this deliverable, final designs of the data governance tools are provided together with the technical achievements of the task.

1.1 Objective of the Deliverable

The objective of this deliverable is twofold: on the one hand, it includes the final design of each of the data governance mechanisms developed within task “T3.5 Data Governance Mechanisms”. On the other hand, it presents the technical advances and final implementation decisions about the development of each tool. Specifically, the governance mechanisms included are the following:

1. A **pseudonymization tool** to pseudonymize enriched transactional data.
2. A **tool for anonymizing datasets** that automatically determines the best anonymization configuration for a particular dataset.
3. A **mobile digital user onboarding service** which uses virtual IDs derived from government issued documents (such as eID cards or passports).

The three tools presented are of significant importance for banks and FinTechs, ensuring that companies or other entities are properly handling personal data or sensitive information. Personal data from insurance companies or financial institutions needs to be anonymized or pseudonymized to protect its privacy, and users of financial services need a reliable way of authenticating their customers using government issued IDs. The three aforementioned tools are demonstrated in real-life environments through the INFINITECH pilots, and are available for external stakeholders at the INFINITECH Marketplace (see [Table 1](#)).

Data Governance Framework	Available at	Pilot
Data pseudonymization tool (JSI)	INFINITECH Marketplace	Pilot #8
Data anonymization tool (GRAD)	INFINITECH Marketplace	Pilots #11, #12
Mobile Digital User Onboarding Service (ATOS)	To be published at Marketplace	Pilot #4

Table 1 - Availability of the Data Governance Frameworks and reference Pilot

1.2 Updates to the previous version

The design of the data governance mechanisms was presented in the D3.13, where the main API design, the architecture decisions, and the workflows of each of the tools are included. In this version, **Section 2 presents the final design of each of the tools** with 1) details about the final architecture and conceptual diagram of the pseudonymization service; 2) the data anonymization tool’s design was updated to include a description of the different operation modes and applicability to different scenarios; and finally, 3) regarding the Digital User Onboarding Service, the functionalities and use cases initially presented in D3.13 are updated and described in detail.

In addition, in section 3, the **final implementation of each of the tools is presented** displaying the functionalities implemented by each of the tools, including details of the supported functionalities (REST API designs, supported metrics or anonymization mechanisms, etc.), or the different interfaces of the digital user onboarding service. At last, we summarise the **Conclusions** gathered from the work in Task T3.5, focusing on the grade of achievement of the objectives and TRL level [4].

1.3 Insights from other Tasks and Deliverables

The work presented is based on the corresponding task “T3.5 Data Governance Mechanisms”, which is included in the “WP3 BigData/IoT Data Management for SHARP Services”. More specifically, this deliverable is related to the following ones:

- **D2.6 Specifications of INFINITECH Technologies - II:** this deliverable exposes the fundamental steps of the process that leads to the specification of the different project technologies that constitute the building blocks utilised by the different project pilots. It includes Input and Output formats, functionalities, and specifications about the implemented technologies (e.g., Big Data/IoT platforms, AI/ML toolkits, HPC infrastructures) that are used to realise them. Within the proposed component groups, there is one related to Security and Privacy including Data Anonymization, Digital User Onboarding System (DUOS) and Data Protection Orchestrator (DPO).
- **D2.8 Security and Regulatory Compliance Specifications - II:** this deliverable specifies the standard and regulatory environment that affects the INFINITECH project. An important part of the document is focused on the GDPR, given its high relevance in BigData and analytics frameworks like INFINITECH. In addition, regulations such as PSD II, MiFiD II and 4AMDL are explored since they are relevant to the financial sector. This set of regulations is analysed and its applicability is assessed with respect to the INFINITECH pilot scenarios.
- **D2.15 Reference Architecture - III:** this deliverable presents the final version of the INFINITECH Reference Architecture (RA). The RA provides a schema to build communication workflows between the different building blocks of the INFINITECH project (as defined on D2.6 Specification of INFINITECH Technologies II), including the Data Governance Mechanisms defined in this document.
- **D3.16 Regulatory Compliance Tools - II:** analyses different regulatory compliance tools and the regulatory requirements of every pilot within the INFINITECH project. The main regulatory compliance tool developed within T3.6 (namely the Data Protection Orchestrator) communicates with the data anonymization component developed within this task.
- **D3.13 Data Governance Frameworks and Tools II:** updates the data governance mechanisms that have been developed within the INFINITECH project.
- **D6.5. Tools and Techniques for Tailored Sandboxes and Management of Datasets - II:** this deliverable describes the tools and techniques to create tailored sandboxes within the INFINITECH project, and describes the “INFINITECH way”, a mechanism for integrated management of testbed based on a continuous integration and delivery approach.

To validate the different mechanisms of data governance, different pilots described as Large-Scale Pilots using the SHARP (Smart, Holistic, Authonomy, Regulatory Compliance and Personalised) model from “WP7 Large-Scale Pilots of SHARP Financial and Insurance Services” will be used, demonstrating the use of the Data Governance Mechanisms in different Financial and Insurance Services scenarios:

- **Pilot #4 “Personalized Portfolio Management (“Why Private Banking cannot be for everyone?”)** uses the Digital User Onboarding Service developed by ATOS for user enrolment and authentication.
- **Pilot #8 “Platform for Anti Money Laundering Supervision (PAMLS)”** demonstrates the implementation of the data pseudonymization mechanisms developed by JSI to pseudonymize enriched transactional data.

- **Pilot #11 “Personalized insurance products based on IoT connected vehicles”** will make use of the data anonymization tool, in particular, the prototype implementation of the location data anonymization mechanisms described in Section 3.2 to anonymize location data reported by connected cars in real time.
- **Pilot #12 “Real world data for novel health insurance products”** will use the data anonymization tool to anonymize data collected in the Healthentia platform.

1.4 Structure

The structure of this document is as follows. Section 1 contains the introduction of the document, the objectives and the relationship with other tasks and deliverables. Section 2 provides the updated design considerations and specifications of the different data governance mechanisms which were developed within T3.5. Section 3 is a technical overview of the final implementations of each tool, together with the technical advancements from the last deliverable. Finally, Section 4 concludes the document summarising the most important concepts and analysing the grade of achievement of the task objectives.

2 Final design of the Data Governance Framework and tools

This section is intended to describe the final design of the data governance framework and tools which were initially presented on D3.12 “Data Governance Framework and Tools I” [2] and D3.13 “Data Governance Framework and Tools II” [3]. In each subsection, a detailed design and the update of each tool is provided, including the requirements, final implementation decisions, or the configuration processes and user interactions, considering the final advancements and implementation results.

2.1 Data pseudonymization tool

Pseudonymization is a data management technology where personal identifiable data fields in a dataset or data record **are interchanged by a set of artificially generated pseudonyms**. By doing this, the data records become less identifiable, but can still be used for data analysis tasks. It is important to highlight that pseudonymization is a reversible process, since the original dataset can be restored to nearly its original state by using the information for re-identification of individuals (for example, translation tables or reversible hashes).

Pseudonymization is defined under the General Data Protection Regulation (GDPR) [1] in Article 4(5). By this definition, **pseudonymized data cannot be attributed to a specific data subject without the use of (separate) additional information**. A single data record might be completely unidentifiable, however, when it is placed into a larger dataset, some information might become identifiable by comparing multiple records. For example, it would be possible to identify the biggest companies by the sum of their transactions in correlation with the timestamps of transactions.

The tool developed within INFINITECH will be used for pseudonymization of financial transactions’ data. The main goal of the pseudonymization tool is to ensure that sensitive data is hidden both during the development and its production use, while still ensuring that the underlying structure of the data is preserved, supporting analysis and reasoning over the pseudonymized data.

The tool supports AES 128, 192, 256 (counter mode only), blowfish, sha1, sha256, sha3, md5 and ripemd160 algorithms. **Dedicated data cleaning procedures were implemented for IBAN and BIC data types**. In addition, the tool also supports noise addition operations, with rounding for numeric and timestamp values.

The diagram on Figure 1 provides a conceptual diagram of the integration of the pseudonymization tool to pseudonymize bank transactions, as in the case of Pilot #8. As depicted in the figure, the pseudonymization service allows users to pseudonymize different datasets using a **common configuration**. The different datasets can be merged under common pseudonyms (new identifiers), allowing data enrichment from different data sources, but preserving anonymity and data privacy. A detailed description of the pseudonymization tool can be found in D3.17 “Regulatory Compliance Tools III” [5].

TRL level of the pseudonymization tool was assessed and demonstrated in a pilot-based environment for the purpose of data enrichment, Pilot #8 “Platform for Anti Money Laundering Supervision (PAMLS)”. External stakeholder publicly available data (from the office for prevention of money laundering) was independently pseudonymized, and noise was introduced in transaction’s data (transaction amounts, dates, etc). The (already) pseudonymized internal *target2 transaction* data (pseudonymized with the same configuration parameters and salt) was enriched with newly acquired external data. The resulting analysis showed that the pseudonymization tool pipeline was successful in both combining the external and internal datasets and masking the private data. Further exploration of enriched datasets showed that additional data (although pseudonymized and noise induced) provided much appreciated insight and information that

was unavailable in the internal dataset, while still easily interpretable in full picture. After the analysis, the **TRL level was assessed as 7**.

The pseudonymization tool is available on the INFINITECH project repository, in the form of a pre-built docker image. The repository also includes documentation and examples. The image is also available in the INFINITECH Marketplace.

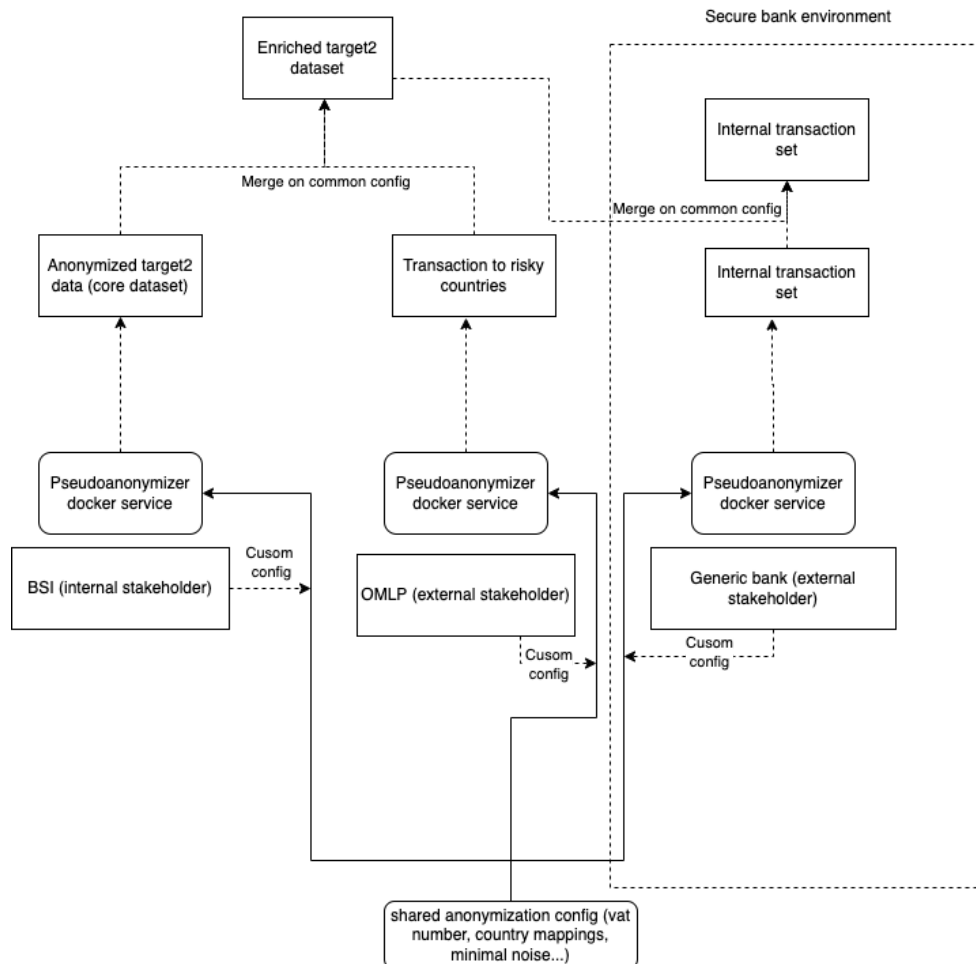


Figure 1 - pseudonymization tool conceptual diagram

The pseudonymization tool will be **exclusively used through a REST API** and therefore, no user interface is envisioned for this. Its flow will be set up with an HTTP POST request, returning a flow key, and the pseudonymization process will be then performed through other HTTP POST requests, using a flow key and the data to be anonymized. **The formal API definition and configuration file description is provided on Section 3.**

There were no specific technical updates during the last iteration from the release of D3.13: the main improvements from the previous version were validation of the implementation in Pilot #8, and minor fixes related to the deployment on the BOS testbed. In addition, during this period the tool was tested successfully on external data.

2.2 Data anonymization

The aim of the anonymization tool is to provide an automated tool to **process personal data to irreversibly prevent identification**. To do so, the data can be modified in many ways and degrees, which will affect the privacy and utility of the resultant data. In general, as the anonymization of the data increases, so does their

privacy, but at the expense of a decrease of the utility or quality. Therefore, there exists a trade-off between these two levels that must be decided by the data owner.

Figuring out how much utility we want to sacrifice to fulfil the data protection requirements of our users is not trivial: the set of anonymization operations that can be applied to a particular dataset is generally high, and measuring the impact on the privacy and the utility is a complicated task. The anonymization tool that is being developed within the INFINITECH project will allow the user to select the anonymization level that best fits its privacy and utility needs, according to a diverse set of metrics.

The last version of the design of the anonymization tool was presented on D3.13 “Data Governance Framework and Tools - II” [3]. This version includes a **Command Line Interface (CLI)** tool to generate and apply the set of anonymization configurations from the desired anonymization operations that the data owner would like to apply to a particular dataset. In this latest version, the CLI tool is still present since it will allow the user to use the tool in environments where graphical user interface is not available. However, the goal of the anonymization tool is to be usable through a **REST API**, that can be invoked directly or using a web application.

In this release we updated the core of the anonymization tool to make it possible to use **Dask Dataframes**³. Dask is a distributed computing library for Python that allows to orchestrate and distribute the computation of large datasets across different working or computing nodes. This library allows the anonymization tool to work with datasets that do not fit in memory. This is a core requirement to fulfil since the anonymization tool is intended to be used in Big Data scenarios. The starting **TRL level** of the anonymization component was 5, and we consider that we will achieve a **TRL level of 7 (technology demonstrated in relevant environment)** at the end of T7.5 “Personalized Usage-Based Insurance Pilots”, where the anonymization tool will be demonstrated in Pilot #11 and Pilot #12.

In this section, the previous version of the document will be updated to provide the **final design of the anonymization tool** for both the CLI and REST API mode.

2.2.1 Operation modes

As mentioned above, the anonymization tool is designed to be used in two different modes: CLI or through a REST API.

CLI mode works as a standalone tool, and has **no external dependencies**. It therefore requires the databases or dataset files to be accessible from the same machine. In addition, it makes use of a **local Dask Cluster**: this means that the Dask distributed computing environment is initiated locally, i.e., in the same machine that is running the anonymization tool. Therefore, it **relies on local machine resources**, but it is still able to distribute the computing of the anonymization operations and metrics across different threads.

REST API mode is more flexible and provides a series of advantages. First, it can use a **remote distributed Dask cluster** that can be deployed locally or on a remote server. The computation of the data will be orchestrated automatically across different workers, allowing for faster processing of large datasets that do not fit in memory. Second, it allows to **execute multiple analysis or anonymization tasks in parallel**, since each operation is performed asynchronously: the analysis and anonymization operations return a task identifier that can be used to track the status of each task. Once the task is completed, the same endpoint returns the results of the operation.

The advanced functionalities of the REST API mode imply a more complicated deployment given the dependencies with external components. These difficulties can be avoided by **following the INFINITECH way [6]**: each sub-component or service can be deployed independently using **docker containers**. Then, we also make use of the **Continuous Integration / Continuous Deployment (CICD)** methodology, and leverage the use of Jenkins to run a set of unitary and integration tests, and to automatically compile the last version

³ Dask: Scalable analytics in Python <https://dask.org/>

of each service into a docker image. This docker image can be pulled automatically by the deployment server (in one of the INFINITECH testbeds). More details of the CI/CD approach, and how the anonymization tool leverages it in each INFINITECH Pilot can be found on D6.2 Testbeds Status and Upgrades - II [7].

2.2.1.1 Command Line Interface

The CLI mode includes all the modules (configuration, analysis, and anonymization) required to perform a successful anonymization of a dataset. As mentioned above, the CLI tool allows the user to execute the anonymization tool in environments where a Graphical User Interface is not available: for instance, a common use case is to need to anonymize a dataset in a remote virtual machine, since personal non-anonymized data cannot be extracted from a company’s premises without being anonymized. In this case, the anonymization tool can be easily deployed in the company’s premises and anonymize the data on-site.

As introduced in D3.13 [3], we designed a **configuration module** to help the user set the different anonymization operations that can be performed over the dataset. [Figure 2](#) shows an example execution of the CLI tool in configuration mode. The tool asks the user to provide the location of the data (1), the delimiter of the data in the CSV file (2), the guessed data types (3), asking the user for confirmation (4), the desired anonymization operations (5), and the set of privacy (6) and utility (7) metrics to be computed.

```

$ python cli.py --mode config --output output.json
Please give the csv file path containing the data []: tests/data/test_1000.csv 1
Please give the csv file delimiter []: , 2
{'dni': 'str', 'date': 'str', 'locality': 'str', 'height': 'int', 'weight': 'float', 'y': 'bool'} 3
Are the types correct? [y/N]: y 4
['categories', 'date', 'delete', 'kmeans', 'same', 'geo_ind']
Please select the operation. Leave blank to finish []: date 5
['dni', 'date', 'locality', 'height', 'weight', 'y']
Please select the fields to perform the operation. Write the fields separated with commas []: date
Please introduce the date format (eg. YYYY-MM-DD / DD/MM/YYYY) []: YYYY-MM-DD
Please introduce the anonymization parameters for the year . Available values are 'same', 'delete' or numeric
value (eg. {year : 10} anonymizes by decade) []: 10
Please select the operation. Leave blank to finish []:
Available metrics: ['CAK', 'R', 'P', 'RCAK'] 6
Please select the privacy metrics. Leave blank to finish []: CAK
['dni', 'date', 'locality', 'height', 'weight', 'y']
Please select the fields to calculate the CAK metric separated by commas []: date, locality, height, weight
Please select the privacy metrics. Leave blank to finish []:
Available metrics: ['AUC', 'MSE', 'MAE', 'MV', 'ILM'] 7
Please select the utility metrics. Leave blank to finish []: MSE
Please select the fields to calculate the MSE metric separated by commas []: date, locality, height, weight

```

Figure 2 – example execution of the CLI Configuration Tool

At the end of the process, the generated configuration (from now on referred to as *analysis configuration*) includes the required information to read the source data, perform the set of anonymization operations, and compute the required metrics. This configuration file can be used in the **analysis mode**. As explained on Section 2.2. of D3.12 “Data Governance Framework and Tools I” [2], this process usually takes place over a representative sample of the data, since it is very time consuming.

The component first verifies the path to the data or the connection to the database where the analysis data is stored, and verifies that the operations are valid. Once the verification is completed, it computes the required privacy and utility metrics and plots the results (see [Figure 2](#)) to allow the user to decide on which anonymization configuration is more suitable for its needs, providing the required trade-off between privacy and utility. Once the process is finished, a JSON file is generated whose content is all the possible

anonymization operations that the tool can perform over the dataset (*anonymization working point*), together with the values of the metrics for each operation.

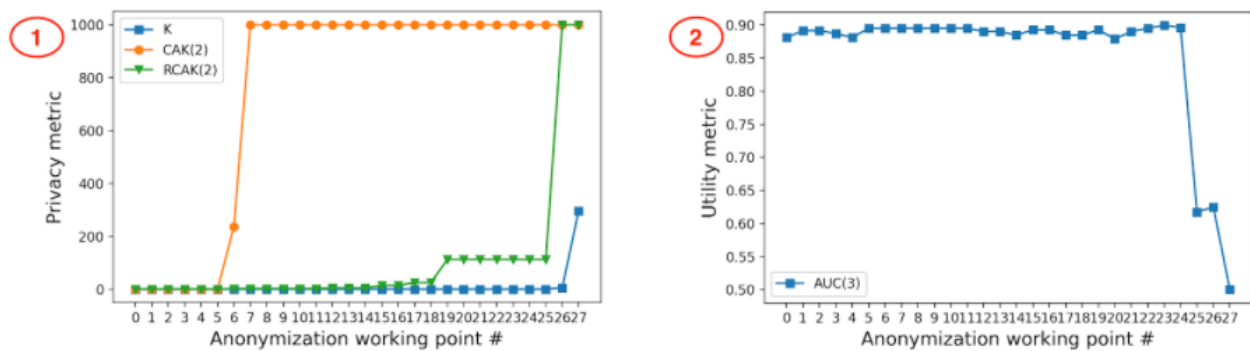


Figure 3 – metric’s results for each configured anonymization operation

At this point, the data owner must decide on **whether one of the provided configurations fits its privacy and utility needs**. [Figure 3](#) shows how the privacy and utility metrics evolve as we apply more aggressive anonymization operations (1), and how the utility evolves towards the applied anonymization (2). In this example, it is possible to see how anonymization configuration #24 provides a good trade-off between privacy and utility, and could be a suitable candidate for the anonymization phase.

Once this first analysis is completed, the user can then apply the selected configuration to the full database or **set a threshold on the value of the privacy and/or utility metric** (for example, utility metric above 0.60). Thus, the anonymization tool will automatically select the configuration that better fits the user needs.

Once the user is ready to perform the data anonymization to the full dataset, the **anonymize mode** will ask the user for the required information to anonymize the final dataset, such as source and destination database locations, and the preferred anonymization preferences (in terms of direct anonymization configuration or specific metri thresholds). Once the process is completed, the tool returns a file with the achieved metric values for the selected anonymization working point.

2.2.1.2 REST API mode

The **REST API** mode is meant to be deployed as a service, and makes use of a remote distributed Dask cluster to perform anonymization operations over a large dataset. This mode supports the execution of multiple anonymization tasks in parallel by making use of an asynchronous API. This can be possible since the analysis and anonymization operations return a task identifier that can be used to track the status of each task. Once the task is completed, the same endpoint returns the results of the operation.

In addition to a set of **anonymization endpoints**, we provide a set of endpoints to **manage configurations and user access**. All the anonymization endpoints require an authentication token that can be only obtained by a valid authenticated user.

POST /user Create a new user in the platform ▼

GET /user/{id} Obtains information about a particular user. Requires admin privileges ▼ 🔒

First, the tool offers an endpoint to **create a new user in the platform**. It receives the required data to create a new user (name, email address, password), and stores the user on an internal database. The created user can be accessed by **administrator users** using its ID.

POST `/login` Login user in the platform



`/login` receives a user name and an encrypted password, and checks if it corresponds to an existing user previously created. If so, it returns a new bearer authentication token to be used on all the API calls during a session. To support userless authentication with other components, the tool will also provide authentication by API Keys (to be implemented in the scope of Task T7.5 for integration with other INFINITECH components in Pilot #11 and Pilot #12).

In addition to **users**, the tool also manages **analysis and anonymization configurations**. This eases integration with other components, since users do not require to store the configuration files themselves, and can access them through the API.

POST `/configuration` Inserts new configuration. Requires authenticated user.



GET `/configuration` Obtains information about user last active configuration. Requires authenticated user.



GET `/configuration/{uuid}` Obtains information about a particular configuration. Requires authenticated user.



PUT `/configuration/{uuid}` Updates particular existing configuration. Requires authenticated user.



DELETE `/configuration/{uuid}` Deletes a particular configuration. Requires authenticated user.



At last, the tool provides a set of **anonymization endpoints** to perform analysis or anonymization operations:

POST `/analyze` Analysis operation



POST `/analyze/{uuid}` Analysis operation with saved configuration



The endpoint `/analyze` performs the configured anonymization operations and privacy/utility metrics calculations. The configuration file can be received directly in a JSON object, or by providing the configuration ID previously stored in the platform by the user. If the configuration is correct, the API **returns an identifier of the analysis task**. This identifier will allow the client to track the progress of the analysis job using the `/progress` endpoint.

POST `/anonymize` Anonymize operation



POST `/anonymize/{uuid}` Anonymization operation with saved configuration



In a similar way, `/anonymize` receives the anonymization configuration, and starts the data processing, storing the resultant database in the destination database.

POST `/guess_types` Returns the identified types of the database variables or columns. You can use this api call to fill `field_types` parameter on `analysis/anonymization` API call

The endpoint `/guess_types` allow to perform a type identification task over a particular dataset. It returns the identified data type of each of the columns, and a sample of ten random rows of the dataset.

GET `/progress/{task_id}` Get the status of a task.

At last, the endpoint `/progress` endpoint receives as a parameter a task identifier, and returns the status of the task, allowing to track the progress of large anonymization or analysis tasks.

2.3 Digital User Onboarding Tool

Digital User Onboarding System (DUOS) is a **solution for dealing with virtual identities in a mobile device**. DUOS can be used in the financial sector of the INFINITECH project to allow bank customers or FinTechs to perform remote registration using electronic ID (eID) cards or passports. DUOS allows them to create their own user identities (virtual eIDs) that will enable them to access the bank or fintech services, but without sharing with them the user's biometric information.

DUOS allows **remote user registration using eID or electronic passport**, providing various identity proofing and verification services in Android mobile devices. These services link the new identity (virtual eID) with a government issued eID or passport. DUOS requires the use of machine readable documents (eIDs issued by European National authorities) according to the EU eID schemas. These documents are standardised by the ICAO Document 9303 (endorsed by the International Organization for Standardization and the International Electrotechnical Commission as ISO/IEC 7501-1) and have a special Machine Readable Zone (MRZ). The MRZ is specially designed to be read from electronic devices, and is usually at the bottom of the identity page at the beginning of a passport. Examples of ICAO compliant documents are eMRTD (i.e., ePassport) and the Spanish official eID card (electronic DNI) [8].

DUOS provides multiple features to ensure the security of the created virtual eIDs:

- **Verification of the electronic data** stored on the chip.
- **Verification of the Machine Readable Zone (MRZ)** of the document using Optical Character Recognition (OCR)
- **Flexible multi-factor authentication** for different users or identities by combining face images captured from the user, with the public key certificates stored in the eID/passport.
- **Integration** with different **Identity and Access Management (IAM) systems**.

Formally, we can identify the following **actors** within DUOS platform:

- **User**: the user is the person interacting with the service which is requesting authorization to obtain a virtual identity. The issued virtual identity is used to authenticate the user and enable access to the service. The user is the **data subject providing personal data** and for which the privacy protection will be implemented.
- **Verifier**. The verifier validates the requests for authentication. The service provider grants access to its services relying on this verification procedure.
- **Service Provider (Relying Party)**: provides the services the user wants to access. For example, in the case of INFINITECH Pilot #4, the Service Provider (SP) would be the portfolio services from Prive.

DUOS supports two different use cases: **enrolment of a new user** (obtaining a new virtual ID, see [Figure 4](#)), and **authentication using an existing virtual ID** ([Figure 5](#)). The enrolment process assumes that the eID has already been issued by an authorised authority, and **only the enrolment of a new virtual identity (vID) falls within the scope of DUOS**. The first step is to read the official identity documents’ MRZ, and extract the data from the chip. Once the data is extracted, the identity of the eID owner is verified using biometric verification using a face recognition process using a mobile device. At last, if the verification process is correct, the virtual ID is issued and stored within the mobile device to be used for authentication with third party entities (banks or FinTechs).

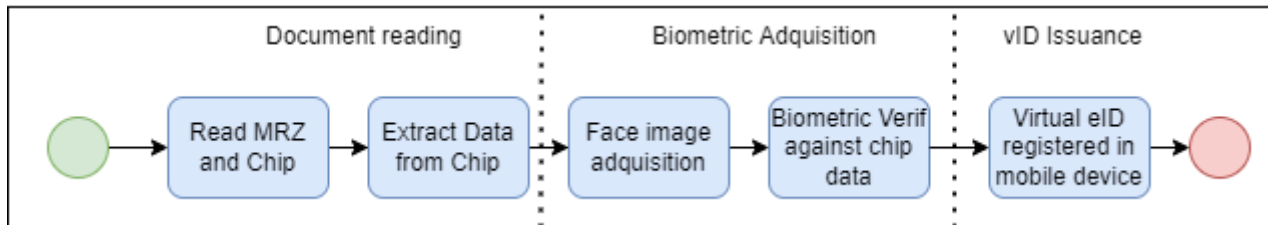


Figure 4 - DUOS enrolment use case

Once the virtual identity is issued, it can be used for **authentication**. The mobile device can store multiple vIDs, so the first step is to select the vID to be used. DUOS will know which third party must authenticate against by reading a QR code on the third party application.

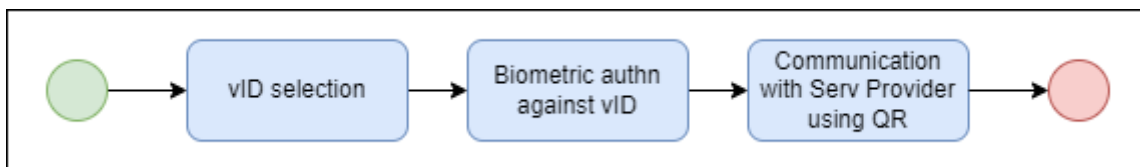


Figure 5 - DUOS authentication use case

The identity of the user is verified (to ensure that the user is who claims to be) by using **biometric authentication**. It is worth mentioning that the biometric authentication process happens **within the mobile device**, and the user’s biometric is not shared with the third party application. In case of a successful authentication, DUOS can communicate with a Service Provider (SP) by reading a QR code that the SP issues. The QR would contain the URL of a Service from the SP. DUOS would use this URL to send the data requested by the service or if the URL is an authentication service of the SP, DUOS would send information to the service to authenticate the user.

Pilot #4 “Personalized Portfolio Management (“Why Private Banking cannot be for everyone?”) has considered the integration of DUOS within its architecture. However, during Pilot development the decision was to not integrate DUOS in the final Pilot architecture. Anyway, DUOS has evolved according to the needs of Pilot #4, both at design level and implemented functionalities, achieving a **TRL level of 6 (system prototype demonstration in operational environment) - from a starting level of 5**: several refinements have been performed on both flows (creation and use of virtual identities) as well as the communication with external services through QR codes, in particular with financial services in Pilot #4.

3 Description of final implementation of data governance tools

This section describes the technical advances during the last period, and provides the final technical specification of each of the tools. In each subsection a detailed technical description of the tools is provided, focusing on the advances achieved from the submission of D3.13 “Data Governance Frameworks and Tools II” [3].

3.1 Data pseudonymization service

As fully detailed in D3.17 Regulatory Compliance Tools III [5], the pseudonymization service is provided as an independent and self-sufficient docker container that can be run independently by stakeholders. All interactions with the service are performed through an stateless REST API (the pseudonymization tool does not save any data, but can be configured to provide a revision trace).

This section provides the detailed REST API specification of the pseudonymization service.

3.1.1 REST API specification

The main pseudonymization functionality is provided by the `/pseudonymize` endpoint. The *flow file* (see [Appendix C: Pseudonymization configuration](#)) describes the required configuration to perform a pseudonymization task. It is a JSON array where each element corresponds to a pseudonymization operation to be performed by the service. The specification of the flow file is as follows, with different required parameters depending on the pseudonymization *type* to execute:

- field: header or key name of fields to process
- type: data type:
 - IBAN
 - BIC
 - text (default)
 - numeric
 - timestamp
- pseudonymization method (case insensitive):
 - bcrypt
 - scrypt
 - aes
 - blowfish
 - sha1
 - sha256
 - sha3
 - md5 (default)
 - ripemd160
- key (required for **aes** and **blowfish** pseudonymization). AES key must be exactly 128, 192 or 256 bits long.
- salt: if not provided it will be randomly generated at initialization
- round_interval (numeric and timestamp fields): The unit for timestamp rounding is days.
- noise_std: standard deviation for gaussian distribution which is used for sampling of the pseudonymized number.

- `format`: Format⁴ of the date in timestamps. Default format: "M/d/yyyy"
- `keylen`: Key length for **scrypt** algorithm
- `cost`: CPU/memory cost parameter for **scrypt**. Must be a power of two greater than one. Default: 16384
- `block_size`: Block size parameter for **scrypt**. Default: 8
- `parallelization`: Parallelization parameter for **scrypt**. Default: 1
- `salt_rounds`: For **bcrypt** algorithm. If salt is not provided a new one is generated using a specified number of rounds.
- `Data file`: data file must conform to specification in flow diagram and can be presented as JSON or CSV format. In the case of CSV data, the first row is a header and other rows are the data itself (the delimiter is automatically inferred). In the case of JSON data, each element must be an object with the field name as key and its value as value.

IBAN and BIC data types are also checked and verified to enable quick feedback loops. IBAN fields will have the following characters removed: " ", "\t", "-", "EUR". IBAN and BIC fields that are blank are not encrypted and are left as is. Any fields that are of BIC or IBAN type will also have its value verified and returned in a new column with a new name `[field]_IBAN` or `[field]_BIC`, where `[field]` is its corresponding field name. Values will be `TRUE` for valid and `FALSE` for invalid. An example configuration file is provided for reference in [Appendix C: Pseudonymization configuration](#).

In case of error, the following errors are reported following the HTTP standard:

- **200 OK** Processed successfully and the data is returned
- **400 Bad Request** Either something is wrong with the flow or data or both
- **432 Bad Flow** Flow is badly specified
- **433 Bad Data** Data is not as expected
- **500 Internal Server Error** Something went wrong during processing of the request

3.2 Data anonymization

As explained on D3.13 “Data Governance Frameworks and Tools II” [3], during the INFINITECH project the **data anonymization tool** was completely refactored to create a unified, modern application, which supports both user interaction through a **CLI tool** and a **REST API**. The underlying processing core of the application is shared by both entry points, leveraging a hexagonal architecture [9]: the data anonymization tool can be driven equally by users, programs, automated tests or scripts, and can be developed and tested in an isolated way. This provides high flexibility during development, and eases the adoption of the INFINITECH way [6] for Continuous Integration and Deployment.

This subsection is intended to serve as a user manual and reference of the supported data anonymization operations, privacy and utility metrics, and the available data stores for data ingestion and storage of anonymized datasets.

3.2.1 Anonymization Operations

For the proper achievement of data anonymization, it is important to carry out a suitable selection of the operations to be used. In this case, the chosen operations are going to be used in both analysis and anonymization service. These operations are configured by two of the fields included in the configuration file (see [Appendix B: Anonymization configuration](#)): `type` (operation name), `fields` (data columns to apply the anonymization operation) and `params` (operation parameters, if applicable).

In the version presented in D3.13 [3], the tool supported the anonymization of text strings, basic dates, and numerical data. This functionality has been extended in the current version to also support the

⁴ <https://github.com/datejs/Datejs/wiki/Format-Specifiers>

anonymization of other types of data such as time, datetime or even GPS positions. The operations considered for carrying out said anonymization together with their respective parameters are presented in [Table 1 - Description of supported anonymization operations](#).

Operation	Parameters	Supported Data	Description
delete	-	int, long, float, double, str, boolean	Erase field values. Used with direct identifiers
same	-	int, long, float, double, str, boolean	Do not modify the field value (no anonymization applied)
rounding	dividers	int, float, double	Round values to the nearest integer
kmeans	<List with the numbers of groups to create>	int, float, double	Used at the analysis service. It receives a list with the number of groups that the clustering algorithm has to generate and returns its centroids as a result
	centroids	int, float, double	Used at the anonymization service. It applies to the values the K-means algorithm by replacing each value with its centroid
date	dividers format	date represented by a string (YYYY-MM-DD by default)	Date is anonymized by generalising the days, months and/or years depending on the parameter. It is possible to specify the format of the date (for instance, MM-DD-YYYY for American date format) by using the <i>format</i> parameter
categories	classes: [{ input output }]	int, long, float, double, str, boolean	Anonymise generalising the values by categories. "input": list of subcategories. "output": general category
gaussian	values, sd, mean	int, float, double	Add noise to a variable for changing its values and not to be exact. values: signal-to-noise ratio to calculate a standard distribution with its mean (sd, mean)
geo_ind	values	coordinate represented by a string (lat, long)	Add noise to a location to satisfy the geo-indistinguishability condition. values: specifies the privacy parameter

Table 2 - Description of supported anonymization operations

3.2.2 Privacy and utility metrics

Once the anonymization operations are set, it is the turn to decide which privacy and utility metric are going to be used in the process. These metrics are used in both analysis and anonymization services, and are configured by the fields contained in the configuration file (see [Appendix B: Anonymization configuration](#)): *type* and *params*.

In terms of privacy, the metrics (*type*) that are considered in its use are the following:

- **K**: Used to indicate how many values are identical to each other. In this case, each value in a database is identical to other $k-1$ values. To be calculated, all the columns of such a database are considered. A value of $K=100$, for example, indicates that groups of at least 100 users can be made, being these users indistinguishable from each other. As a *result* of this metric, it is possible to obtain the metric value and as *advanced_results*, the group size that can be formed, as well as the index of a row or element of each group.
- **P**: Represents the re-identification risk in the database, i.e., the probability of identifying an individual in the database. This parameter can be considered as the inverse of K ($1/K$) represented as a percentage. The *result* of this metric is its value.
- **CAK** (Constrained Attacker K): This metric is a less restrictive extension of the K metric. As with the K metric, each of the values in a database is identical to another $k-1$ values, but in this case only considering the columns specified in the param *fields*. Therefore, this metric considers the presence of an attacker with a certain likelihood of knowing the values of the columns in the database (for instance, it might be easier to know the age of a person than its exact weight). The *results* of this metric provides its metric value, meanwhile the *advanced_results* contains the group size that can be formed, the index of a row or element of each group and the fields used to calculate the metric at each working point.
- **RCAK** (Random Constrained Attacker K): This metric can be considered as an extension of the previously mentioned CAK. As the previous one, each of the values in a database is identical to other $k-1$, but with the difference that, in this case, a random subset of columns is considered, defined by the *num_fields* parameter. In its performance, a number of iterations, stated by the parameter *rounds*, are executed, selecting in each of them *num_fields* different columns, and measuring the value of K to finally provide its mean value. This metri tries to take into consideration the presence of an external attacker with a certain likelihood of knowing the values of the columns, although it is not known exactly which of them is the most probable to know (so a random sample of N columns is selected). As a *result*, the value of the metric is obtained, meanwhile the *advanced_results* contain the fields used to calculate the metric in each *working point*.

When it comes to the utility, the metrics used are:

- **MSE**: (Mean Squared Error) Measures the average of the squares of the errors introduced with the anonymization operations. The *result* obtained provides the metric values, i.e. the root Mean Square Error of the specified columns.
- **MAE**: (Mean Absolute Error) Represent the arithmetic average of the absolute errors. It provides the Mean Absolute Error of the specified columns as *result*.
- **MV**: (Mean Variation) Measures the mean variation of the data.
- **ILM**: (Information Loss Measure) An extension of the Mean Error Metric. It scales the values by the square root of two times the standard deviation of the difference between both vectors. As a *result* it provides the Information Loss Measure.
- **distance**: Estimates the mean distance between the original coordinates and the anonymised ones.

3.2.3 Data Sources

The anonymization tool supports data ingestion and storage in multiple formats; the aim of the tool is to be flexible and adaptable to different scenarios: from anonymization of datasets that cannot leave the

premises of a particular company, to integration on complex data workflows. To do so, we support both static datasets (such as files and SQL databases) and anonymization of continuous streams of data.

3.2.3.1 Static datasets

The anonymization tool supports processing of datasets in both CSV and JSON formats. CSV and JSON are the most used file formats to store structured data: while CSV stores the values in a table-like format, JSON is structured as attribute-value pairs and arrays.

Regarding SQL databases, the anonymization tool uses the direct integration of Pandas and Dask Dataframes to read different databases, without the need to explicitly implement the queries. Thanks to this integration, we support any SQL database supported by Pandas; internally, Pandas uses the SQLAlchemy framework, which includes dialects for SQLite, PostgreSQL, MySQL, Oracle, MS-SQL, Firebird, Sybase, and others. In addition to the support for common SQL dialects provided by Pandas, we also added compatibility with the data management layer of INFINITECH, also known as INFINISTORE, and its dual SQL/NoSQL connector developed under the scope of T3.1 and reported in D3.3 “Hybrid Transactional/Analytics Processing for Finance and Insurance Applications - III” [10].

3.2.3.2 Data in streaming

Nowadays, it is common that data does not arrive statically to a system (in a database or static files), but continuously or in streaming. Data processing in streaming adds certain difficulties when it comes to making a prior analysis of the data to select the best anonymisation strategy. Since the data arrives in portions to the system, the information is always incomplete, and performing a correct privacy risk assessment and utility evaluation is not an easy task. Thus, the anonymization tool does not support analysis of datasets in streaming, but it supports the **anonymization of data in a continuous way**; Once the anonymization strategy has been decided using the analysis mode, we support the anonymization of streams of data given a particular anonymization configuration.

To handle the amount of data that can be received in streaming, we leverage the use of Apache Kafka⁵, an open-source event streaming platform. The data to be anonymised in streaming can be sent to the anonymization tool using a POST request (providing the data to be anonymized in the request body). The streaming data module has been designed to abstract end users from the final implementation, being transparent about the queue or data ingestion system used internally.

[Figure 6](#) shows the detailed workflow of an anonymization task with data in streaming: to start anonymizing data in streaming, the end user must first **configure the anonymization operations** to be performed over the data, and indicate that the data will be received in streaming using the */configuration* endpoint, providing an anonymization configuration with data source “streaming” (see [Appendix B: Anonymization configuration](#)). With the first call to */anonymize* with a streaming configuration, the anonymization tool creates a new Kafka topic, and starts an internal batch consumer to process the received data continuously. Subsequent petitions will be now redirected to the assigned topic. The anonymization component **consumes the data from the topic**, applies the corresponding anonymization operations, and stores the resultant data in the destination database.

⁵ Apache Kafka <https://kafka.apache.org/>

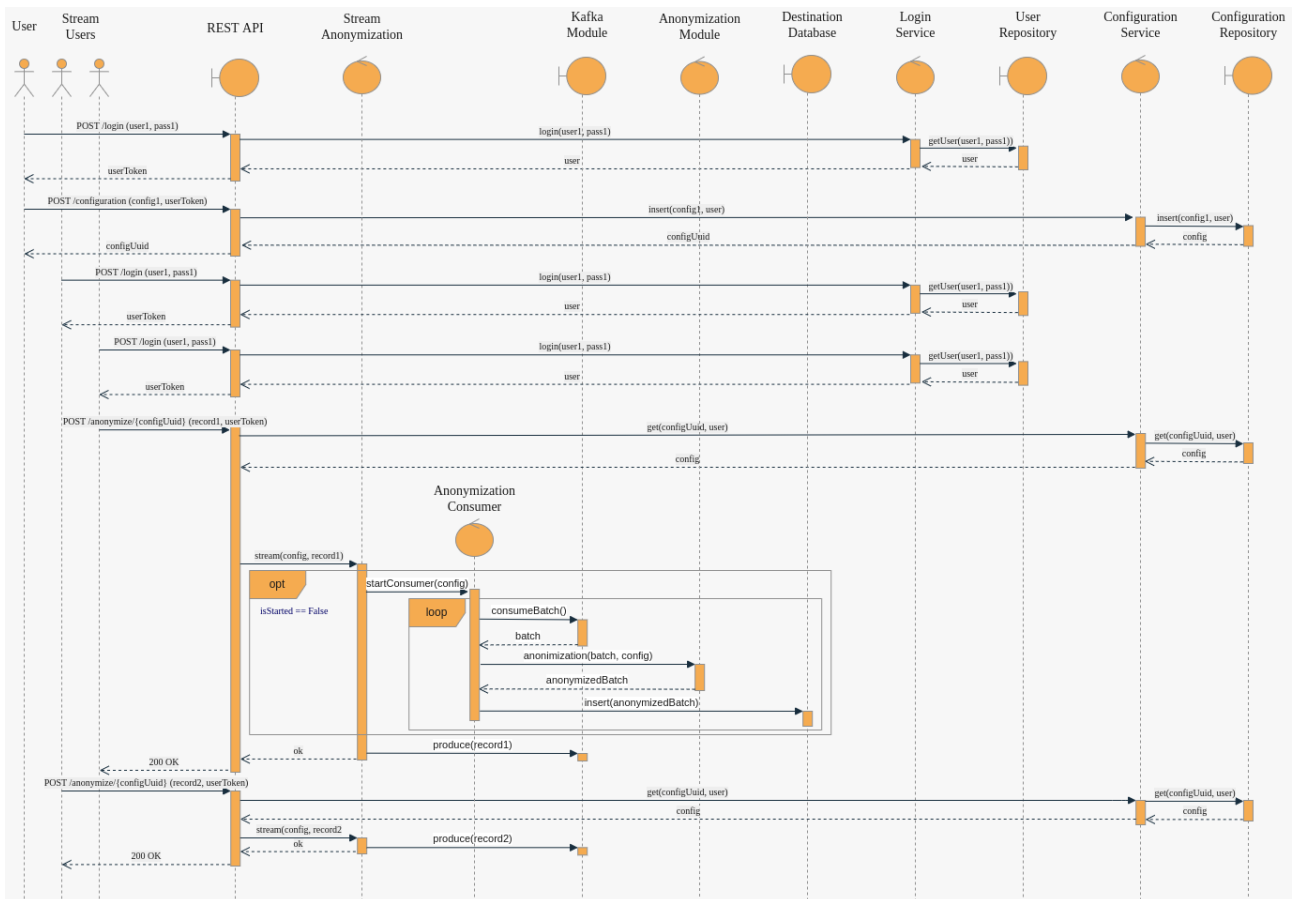


Figure 6 - workflow of data anonymization in streaming

In this way the **users are abstracted from the data stream and topic creation** and all requests have the same format regardless of how many users send data for a streaming anonymization process and in which order they make those requests.

3.3 Digital user onboarding services tool

DUOS implementation within the INFINITECH project is an adaptation of the Digital User Onboarding System developed in the context of ARIES project [11] This section is intended to extend the description of DUOS interfaces and architecture already presented in D3.13 “Data Governance Frameworks and Tools II” [3] by focusing on the flow execution of the DUOS mobile application for Android, already available in the [INFINITECH repository](#) as an APK. The DUOS mobile application will be also available to external stakeholders through the INFINITECH Marketplace.

The first time that the app is used in the mobile phone there are not any virtual identities created in the device. After the initial screen, the only option presented to the user is the creation of a new identity ([Figure 7](#)).

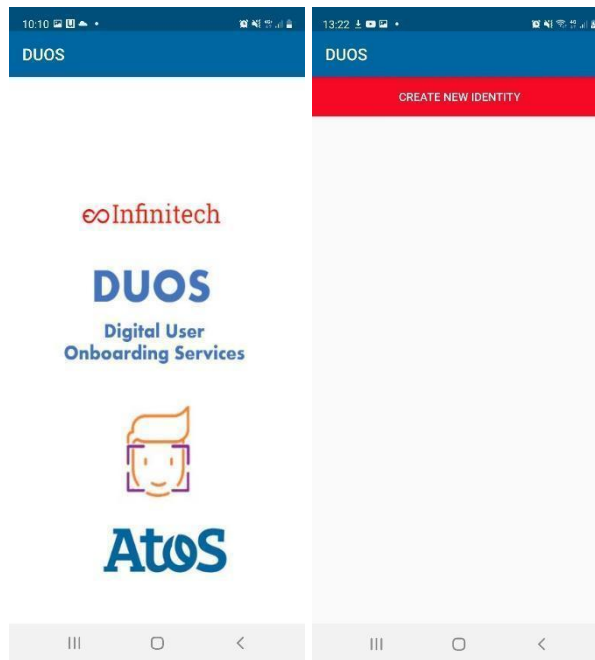


Figure 7 - DUOS app initial screen (left) and new identity creation menu (right)

The creation of a new identity starts by asking the user for consent for using the device’s camera, in order to read the MRZ of a compatible eID card. DUOS detects automatically when the MRZ is presented to the camera by using OCR technology (see [Figure 8](#)). Once a valid MRZ is read, DUOS extracts the associated data (identity number, birth date, and the document expiration date).

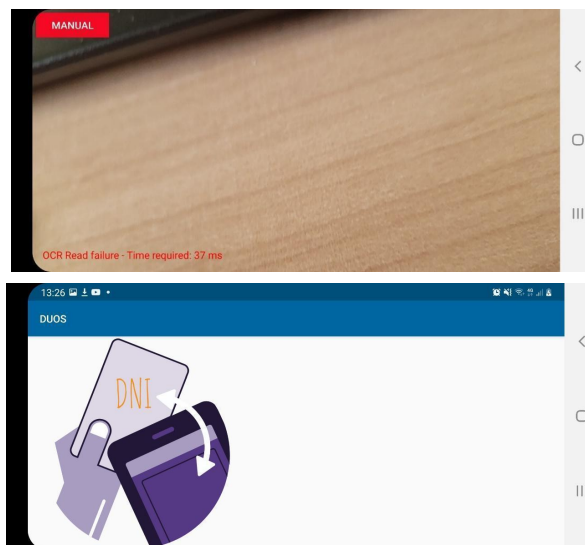


Figure 8 - DUOS waiting for a MRZ zone (left) and DUOS asking for ID reading NFC (right)

Once the data is read, DUOS asks the user to approach the eID to the NFC reader of the mobile device, to read the chip of the electronic document, using the identifier obtained through MRZ reading.

Finally, the data from the chip is obtained (face image, name, last name, document number, birthdate, and Gender). After this step, the virtual identity is complete and can be saved in the application for its use in an authentication procedure, as explained in Section 2.3.

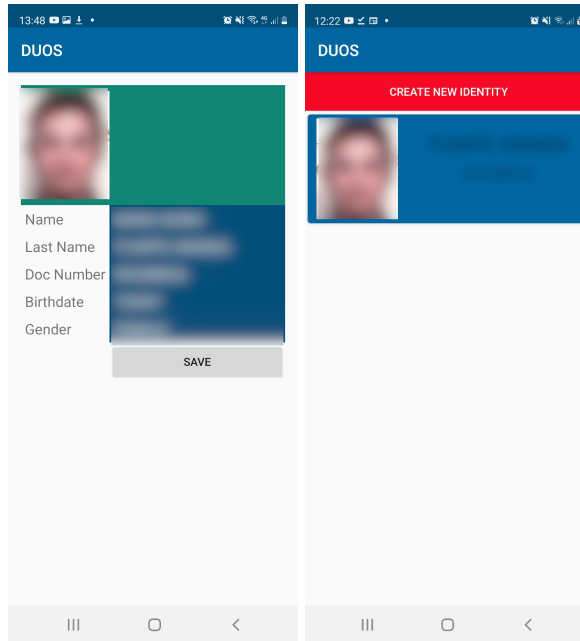


Figure 9 - DUOS virtual identify created (left) and virtual identities available (right)

To use the recently created virtual identity, the first step is to open DUOS and select the desired virtual ID. Once selected, DUOS asks the user to take a picture of the person trying to use the virtual ID, to verify that it is the same person whose picture is stored in the virtual identity. At this point, DUOS starts the face verification process ([Figure 10](#)).

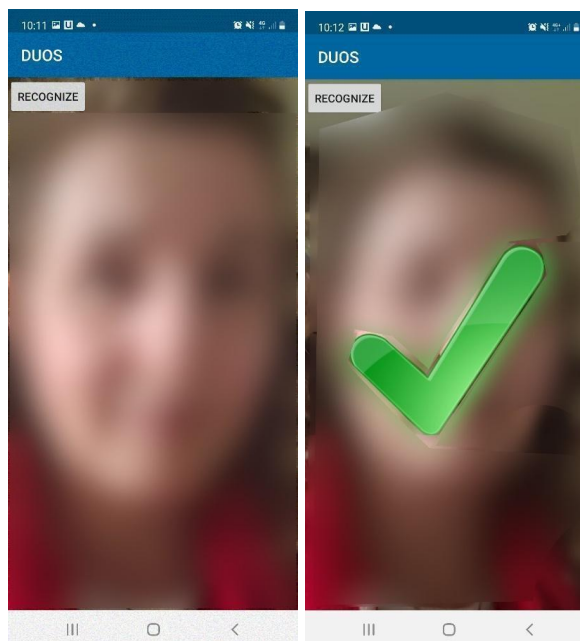


Figure 10 - DUOS face recognition starting (left) and face verification process completed(right)

Once the identity of the virtual identity owner is verified, DUOS allows the user to read the QR code from the external application that the user wants to authenticate against.

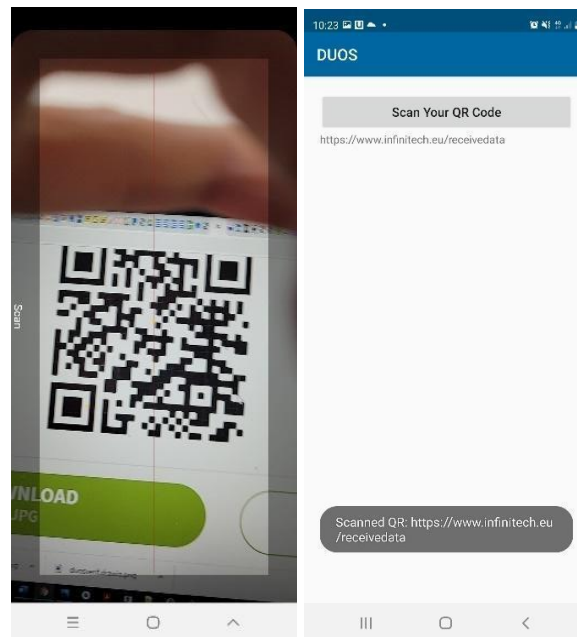


Figure 11 - DUOS scanning a SP QR code (left) and QR scan process completed (right)

Finally, DUOS asks the user if he/she wants to share the requested data with the third party app (Service Provider), as shown in [Figure 11](#). If so, DUOS reads the QR code to obtain the URL to send the required data.

4 Conclusions

This document reports the achievements of the task “T3.5 Data Governance Mechanisms” of the INFINITECH project, whose goal was to implement and provide the following data governance building blocks: (i) a pseudonymization tool; (ii) a mechanism for anonymizing datasets and (iii) a mobile digital user onboarding service with virtual eID derived from government issued documents. Towards this goal, the final design and implemented functionalities of each of the tools was presented, updating the contents of D3.13 “Data Governance Frameworks and Tools II” [3].

The **pseudonymization tool from JSI** includes support for repeatable unique (pseudo)anonymization of identifiers (IBAN numbers, VAT numbers...) with additional generalisation techniques, including support for numeric data and timestamps along with noise induction to prevent deanonymization. The tool provides a simple to use REST API and can be configured with a configuration file. The configuration file can be shared among different stakeholders and thus providing the same anonymization results while keeping the original data secure. The achieved **TRL level** of the pseudonymization tool was assessed as **7 (system prototype demonstration in operational environment)** after its integration on Pilot #8. It is already part of the [INFINITECH Marketplace](#).

The **data anonymization tool** from GRAD helps users to handle personal data to irreversibly prevent identification. Usually, analysis of privacy requirements of a dataset and whether the data meets GDPR requirements is a long, tedious process, and requires certain knowledge about the different anonymization operations that can be applied to a particular dataset. We consider that the tool designed and implemented during T3.5 achieved all its objectives: (i) add new anonymization algorithms, such as the support for GPS data anonymization, (ii) add new privacy and utility metrics to measure the re-identification risk, and (iii) improve the process to determine how the data should be anonymized.

During T3.5, the tool evolved from two isolated modules, which required manual configuration, to a modern, scalable application that can handle large datasets, and can be interfaced through both a CLI module or a REST API. The starting TRL level of the anonymization component was 5, and we consider that we will achieve a **TRL level of 7 (system prototype demonstration in operational environment)** at the end of T7.5 “Personalized Usage-Based Insurance Pilots”, where the anonymization tool will be fully integrated in Pilot #11 and Pilot #12. The data anonymization tool is available both at the INFINITECH repository, and at the [INFINITECH Marketplace](#) as a set of docker containers.

Paying attention to the **Digital User Onboarding Service (DUOS) from ATOS**, it provides a way of enrolling new users into a service in a remote and secure way, by allowing the creation of virtual identities derived from eID cards or passports, combining the electronic certificates stored in the chip with biometric recognition for increased security. During T3.5, DUOS has been improved to fit the needs of Pilot #4. The Pilot initially considered DUOS as a potential use case, but it is not going to be integrated in the final pilot solution. Anyway, DUOS has evolved and improved during T3.5 to include new use cases and functionalities, more specifically, to fit the legal requirements for authentication with Spanish entities (face recognition and scanning of government issued id). The starting TRL level for DUOS was 5 and **it has been improved to 6 (technology demonstrated in relevant environments)**. Several refinements have been performed on both flows (creation and use of virtual identities) as well as the QR way of communication with the external service (financial entity). DUOS is available in the [INFINITECH repository](#) as an APK for Android phones, and will be published in the INFINITECH Marketplace too.

The **final validation** of each of the tools will be included in the **corresponding deliverables from WP7**, that will detail the integration and validation of the Pilot use cases by leveraging the Data Governance Mechanisms developed during T3.5.

Appendix A: Literature

1. European Commission. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
2. “INFINITECH-D3.12 - Data Governance Framework and Tools - I.” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/39>.
3. “INFINITECH-D3.13 - Data Governance Framework and Tools - II.” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/40>.
4. European Commission, “Technology readiness levels (TRL” [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf
5. “INFINITECH-D3.17 - Regulatory Compliance Tools - III” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/44>
6. “INFINITECH-D6.5 - Tools and Techniques for Tailored Sandboxes and Management of Datasets - II.” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/81>
7. “INFINITECH-D6.2 - Testbeds Status and Upgrades - II.” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/78>
8. ICAO, “ICAO Document 9303 - Machine Readable Travel Documents.” <https://www.icao.int/publications/pages/publication.aspx?docnum=9303> (accessed Mar. 16, 2022).
9. Alistair, Cockburn (2005-04-01). "Hexagonal architecture". [Online] Available alistair.cockburn.us.
10. “INFINITECH-D3.3 - Hybrid Transactional/Analytics Processing for Finance and Insurance Applications - III ” [Online]. Available: <https://app.infinitech-h2020.eu/deliverable/30>
11. “ARIES - ReliAble euRopean Identity EcoSystem.” [Online]. Available: <https://www.aries-project.eu/>

Appendix B: Anonymization configuration

```
{
  "source_type": "file",
  "database": {
    "source": {
      "db_type": "csv",
      "path": "tests/data/test_1000.csv",
      "delimiter": ",",
    },
    "destination": {
      "db_type": "csv",
      "path": "tests/data/output_path.csv",
      "delimiter": ",",
    },
    "field_types": {
      "dni": "str",
      "date": "str",
      "locality": "str",
      "height": "float",
      "weight": "float",
      "y": "int"
    }
  },
  "analysis": [
    {
      "type": "kmeans",
      "fields": [
        "kmeans1",
        "kmeans2"
      ],
      "params": {
        "values": [
          2,
          1
        ]
      }
    },
    {
      "type": "categories",
      "fields": [
        "categories3",
        "categories4"
      ],
      "params": {
        "values": [
          0
        ]
      }
    }
  ]
}
```

```
[
  {
    "inputs": [
      "Pontevedra",
      "Vigo"
    ],
    "output": "Pontevedra"
  },
  {
    "inputs": [
      "Santiago",
      "Coruna"
    ],
    "output": "Coruna"
  }
]
},
{
  "type": "delete",
  "fields": [
    "delete5"
  ],
  "params": {
    "values": [
      0
    ]
  }
},
{
  "type": "same",
  "fields": [
    "dependant6"
  ],
  "params": {
    "values": [
      0
    ]
  }
}
],
"metrics": {
  "privacy": [
    {
      "type": "K",
      "params": {
        "values": [
          null
        ]
      }
    }
  ]
}
```

```
},
{
  "type": "CAK",
  "params": {
    "values": [
      "kmeans1",
      "kmeans2"
    ]
  }
},
{
  "type": "RCAK",
  "params": {
    "values": {
      "rounds": 3,
      "num_fields": 3
    }
  }
},
],
"utility" : []
}
}
```


Appendix C: Pseudonymization configuration

```
[
  { "field": "PLACNIK",
    "type": "IBAN",
    "method": "md5"},
  { "field": "PREJEMNIK",
    "type": "IBAN",
    "method": "aes",
    "key": "4428472B4B6250655368566D59713374"},

  { "field": "BANKA_PLACNIKA",
    "type": "BIC",
    "method": "sha3" },

  { "field": "DATUM_PLACILA",
    "type": "timestamp",
    "method": "round",
    "round_interval": 3,
    "noise_std": 15,
    "format": "M/d/yyyy" },

  { "field": "ZNESEK_PLACILA",
    "type": "numeric",
    "method": "round",
    "round_interval": 1000,
    "noise_std": 5000 }
]
Example data:
[
  {
    "PLACNIK": "AL47 2121 1009 0000 0002 3569 8741EUR",
    "PLACNIK_OBD": "6706491",
    "BANKA_PLACNIKA": "",
    "PREJEMNIK": "SI56014019999000054",
    "DATUM_PLACILA": "11/19/2009",
    "ZNESEK_PLACILA": 20022000
  },
  {
    "PLACNIK": "NO 93 8601 1117947",
    "PLACNIK_OBD": "6476619",
    "BANKA_PLACNIKA": "PUADAOLU",
    "PREJEMNIK": "SI56010000003300023",
    "DATUM_PLACILA": "10/6/2015",
    "ZNESEK_PLACILA": 1897849.46236559
  },
]
```

```
{
  "PLACNIK": "SI56-1010-0004-8924-680",
  "PLACNIK_OBD": "303596",
  "BANKA_PLACNIKA": "BLICUYMMXXX",
  "PREJEMNIK": "SI56031601000813381",
  "DATUM_PLACILA": "9/1/2016",
  "ZNESEK_PLACILA": 250207.387096774
}
```