Funded by the European Union
Horizon Europe
(HORIZON-CL4-2021-HUMAN-01-27
AI to fight disinformation)

Ref. Ares(2023)5877058 - 29/08/2023

www.ai4trust.eu

D5.4

AI4TRUST

Platform Specification

PARTNERS

Project acronym	AI4TRUST
Project full title:	AI-based-technologies for trustworthy solutions against disinformation
Grant info:	ID 101070190-AI4TRUST
Funding:	EU-funded under Digital, Industry, and space Overall budget € 5.950.682,50
Version:	1.0
Status	Final version
Dissemination level:	Public
Due date of deliverable:	31 August 2023
Actual submission date:	29 August 2023
Work Package:	WP5
Lead partner for this deliverable:	FINC
Partner(s) contributing:	FBK, CERTH, UNITN, UPB, SAHER
Main author(s):	Marcello Paolo Scipioni (FINC), Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)
Contributor(s) and reviewer(s):	Matteo Saloni (FBK), Vitor H. Bezerra (FBK), Riccardo Gallotti (FBK), Serena Bressan (FBK), Maria Vittoria Zucca (FBK), Andrew Staniforth (SAHER), Juliet Lodge (SAHER), Lampis Apostolidis (CERTH), Symeon Papadopoulos (CERTH), Niculae Sebe (UNITN), Horia Cucu (UPB), Alexandru Caranica (UPB)

Summary of modifications

VERSION	DATE	AUTHOR(S)	SUMMARY OF MAIN CHANGES
0.1	28/06/2023	Marco Giovanelli (FINC), Gabriel H. Carraretto (FINC)	First version of the deliverable
0.2	18/07/2023	Andrew Staniforth (Saher), Juliet Lodge (Saher)	Added SAHER contributions
0.3	27/07/2023	Matteo Saloni (FBK), Horia Cucu (UPB), Alexandru Caranica (UPB)	Added FBK and UPB contributions
0.4	22/08/2023	Lampis Apostolidis (CERTH), Symeon Papadopoulos (CERTH), Niculae Sebe (UNITN)	Added CERTH and UNITN reviews
0.5	25/08/2023	Riccardo Gallotti (FBK), xxx, Serena Bressan (FBK), Maria Vittoria Zucca (FBK)	Added FBK reviews
0.6	28/08/2023	Marcello Paolo Scipioni (FINC), Marco Giovanelli (FINC)	Internal revision and finalisation
1.0	29/08/2023	Serena Bressan (FBK)	Final draft document review for submission to the EC

Table of contents

List of acronyms

ACRONYMS	MEANING
AI	Artificial Intelligence
API	Application Programming Interface
DAG	Directed Acyclic Graph
DoA	Description of Action
IP	Intellectual Property
PS	Platform Specification
SW	Software
UI	User Interface
WP	Work Package

List of figures

List of tables

Executive summary

The AI4TRUST **D5.4 - Platform Specification** is the first technical deliverable (D) of the project and describes the overall aspects of the platform's structure. D5.4 provides the platform's specifications as detailed in Task (T) 5.1 - AI4TRUST Platform Specification of **Work Package (WP) 5 - Technical implementation of the platform & security framework**.

This is the first deliverable of **Work Package (WP) 5** of the European project **AI4TRUST - AI-based-technologies for trustworthy solutions against disinformation**. This deliverable also represents the first part of the means of verification of the **first project Milestone (M1) "AI4TRUST requirements"**, together with the contents foreseen in D6.1 - Pilot Planning Report, i.e., *"Platform and pilot requirements defined. Technological specifications and legal related requirements collected"*.

Specifically, the focus of this document is to describe the design of the AI4TRUST platform, based on an in-depth analysis of the current state-of-the-art solutions, and the legal and ethical implications concerning the detection of online disinformation and misinformation.

The deliverable is structured in five different sections, **introducing the project** and its goals, providing a summary of the **technical requirements** defined in the initial steps of the project, reporting on the initial **platform's architecture definition**, discussing all the problems regarding **ethics**, **security,** and **privacy,** and describing the **further steps** to be taken and how the platform advancement will be handled in the next iterations.

This initial release will focus on the **most relevant components** of the architecture and will be progressively updated to reflect the evolution of the **AI4TRUST platform**.

1. Introduction

The AI4TRUST project aims to establish a **hybrid platform** that combines the effectiveness of advanced **artificial intelligence** (**AI) solutions** with the expertise of fact checkers and journalists (hereafter also 'media professionals') **to fight mis/disinformation,** supporting media professionals and policy makers.

The AI4TRUST platform will operate in **near real-time**, monitoring various online social platforms and news data sources (e.g., web news feeds, news aggregators), effectively filtering out irrelevant information and analysing **multimodal content** (text, audio, visual) across multiple languages.

By integrating quantitative indicators about the trustworthiness of a news item, and incorporating advanced approaches from the social and computational sciences, the AI4TRUST platform will provide media professionals with reliable and explainable data analysis components that can be used **to assess the credibility of news items and debunk mis/disinformation**.

The AI4TRUST platform will follow a **human-centred approach** aligned with European values and is expected to become a standard tool for data analysts fighting against disinformation and misinformation.

This deliverable **D5.4 - Platform Specification** of the **AI4TRUST Work Package 5 (WP5) – Technical implementation of the platform & Security Framework** defines the initial platform's structure (sec. 3.1) based on a list of requirements (sec. 2), specifies the mechanisms for the ingestion (sec. 3.2), analysis (sec. 3.3) and manipulation (sec. 3.4 and 3.5) of data, defines how the given data will be accessed (sec. 3.6 and 3.7), and the problems related to ethics, security and privacy (sec. 4).

Deep technical details and component specifications will be left to subsequent iterations (sec. 5), according to the needs and the evolution of the project.

2. Requirements

This section describes the **collected requirements** that guided the design of the AI4TRUST platform. The clear definition of these requirements and their prioritisation is essential to define the features of the platform and drive their development depending on their priority.

The prioritisation of requirements has been carried out according to the so-called **MoSCoW method**[1], a practice well-known in project management. Its name is derived from the initial of the category names used: **M**ust, **S**hould, **Co**uld and **W**on't. The categories are defined in **Table 1**.

CATEGORY	MEANING
Must	The requirement needs to be included for the project to be considered a success
Should	The requirement is important but not strictly necessary for the current release to be considered a success
Could	The requirement is desirable but less critical and can be considered "nice to have"
Won't	The requirement is not planned for delivery

Table 1 *– MoSCoW Categories*

The current list of requirements has been created at this stage to support the needs of the project. Their current formulation is based on: the described solutions in the AI4TRUST project Description of Action (DoA); the discussions, since the beginning of the project, with the technology providers and the fact checkers of AI4TRUST, taking into account the planned characteristics of the platform at the current stage.

The current list of requirements will be **checked and revised in the next iterations**. Checks and revisions will aim to confirm existing requirements or modify them by providing more clarifications and better rewording; split existing requirements into multiple additional requirements; delete or deprecate existing requirements; or create new requirements. In the latter case, no reuse of old numbers will be made in case of deletions, and new identifying numbers will be used for new requirements, to avoid any possible inconsistencies.

In **Table 2** the list of requirements is shown, which will be used to define the platform structure.

[1] https://en.wikipedia.org/wiki/MoSCoW_method

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE[2]
01	The AI4TRUST platform supports near-real-time data collection from the following data sources: Twitter[3], Facebook, YouTube, web news feeds and news aggregators through publicly available APIs.	Must	F
02	The AI4TRUST platform extends near-real-time input from data sources other than Twitter, Facebook, YouTube, web news feeds and news aggregators (e.g., GDELT, news API).	Could	F
03	The data-gathering process is defined by a set of keywords, associated with the different topics that will be taken into account.	Must	NF
05	Data cover at least seven of the most spoken languages in the EU: English, French, German, Greek, Italian, Polish and Spanish, representing 70% of the EU population in terms of first language spoken.	Must	NF
06	Data annotated/labelled from fact checkers' platforms, together with its metadata (e.g., URLs pointing to textual, audio or visual content, content type, classification) are integrated into the consortium internal API.	Must	F
07	New fact checked information produced by the human fact checkers network is periodically added (once a week if manually, unlimited if automated) in the AI4TRUST platform.	Must	F
08	The data gathered from different platforms are harmonised and pre-processed in nearly real-time.	Must	F
09	The use of the data stored in the AI4TRUST platform is granted using a specifically designed API.	Must	F
10	The development adopts a microservice-based approach to assure the dynamic scalability of the AI4TRUST platform.	Must	NF
11	The development adopts a DevOps approach able to speed up the deployment and fully support the project's agile approach.	Must	NF

[2] F = functional; NF = non-functional.
[3] Now called "X".

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE[2]
12	Standards for API formalisation (e.g., OpenAPI - https://www.openapis.org/) are employed to support API implementation in different development contexts, favour the use of automatic code generation tools (e.g., service stubs) and reduce SW bugs.	Must	NF
13	The AI4TRUST platform supports ethical and security guidelines inherent with the methodological options available.	Must	NF
14	The AI4TRUST platform meets the ethical, privacy and data protection requirements, also in terms of AI explainability.	Must	NF
15	The AI4TRUST platform provides specific technical solutions to enhance privacy and data protection according to the current regulation (i.e., GDPR).	Must	NF
16	The AI4TRUST platform supports high-performance near-real-time data processing streams.	Should	F
17	The AI4TRUST platform supports non-real-time data input.	Should	F
18	The AI4TRUST platform supports non-real-time data processing.	Should	F
19	The AI4TRUST platform enables external (re-)training of AI models.	Should	NF
20	The AI4TRUST platform supports connection with external processing services.	Must	F
21	The AI4TRUST platform offers centralised CPU computation for the inference phase.	Must	NF
22	The GPU-based processing is performed only in partners' premises.	Must	NF
23	The egress of RAW data is limited.	Must	NF
24	The AI4TRUST platform supports the storage of large quantities of data.	Should	F

#	REQUIREMENT DESCRIPTION	PRIORITY	TYPE[2]
25	Standardised schemas for data storage are defined.	Must	NF
26	Existing data from fact checkers are imported into the AI4TRUST platform.	Must	F
27	Results from automated and manual processing of data are presented to users through a user interface.	Must	F
28	The AI4TRUST platform provides low-latency access to data that need to be exposed in the UI.	Should	NF
29	The data analysis and classification components are initially trained and validated on a data set curated by humans.	Must	NF
30	The publicly available component of the AI4TRUST platform is an Open Observatory, which will allow users to capitalise on the output of the complex data analytics, relying on human-machine interactions, employing visual and interactive dashboards that can be easily self-customised to fit their needs.	Must	F
31	The AI4TRUST platform supports three groups of users: 1) fact-checkers, journalists and media practitioners; 2) policy makers; 3) researchers.	Must	NF

***Table 2** – Requirements List*

The list of requirements will be **updated periodically** throughout the development of the project, based on the feedback gathered by all partners and on the advancement of the platform implementation.

A risk related to the accessibility of source APIs needs to be addressed: although, as already included in the Grant Agreement, the possibility of access to **YouTube**'s textual data through their YouTube Researcher programme and **Facebook** via the Crowdtangle service are confirmed, at the time of publication of this deliverable, the **Twitter API** that has been available in the past for research purposes is no longer accessible due to a change in the API access policy recently implemented by Twitter. This shift in API accessibility had been anticipated as a potential risk and was duly considered within our project's contingency plan. In response to this predicament, we promptly adopted a proactive stance by initiating **efforts to establish formal collaborations with alternative social media platforms**.

In particular, we have initiated ongoing dialogues with representatives from **Bytedance**, the owners of **TikTok**. These discussions have revolved around the pursuit of access to the TikTok research API, a promising avenue. However, it's important to note that this API is currently exclusively accessible within the U.S. There are expectations for its launch as part of their European program this summer, a development that presents an enticing opportunity. Given TikTok's extensive popularity, especially among younger demographics across Europe, it stands as an ideal platform for scrutinising and unravelling the intricate patterns through which misinformation and disinformation spread. Although the unavailability of Twitter data presents its share of

challenges, the potential insights gleaned from studying TikTok's environment hold the promise of substantial compensation. A second contingency plan is being considered, which also includes the purchase of a so-called 'basic' access right to the Twitter API.

3. Architecture

This section describes the **overall architecture of the AI4TRUST platform**, its main components, their interactions and how the data flows between those components.

3.1. Overview

The adopted platform design and development approach in AI4TRUST aims to combine the strengths of **AI technologies** with the **expertise and critical thinking abilities of humans** in a platform that will operate in near real-time, targeting multiple online social platforms, and producing a series of analyses based on **multimodal-multilanguage content**. **Figure 1** shows a high-level representation of the platform's structure, based on the requirements listed in **Table 2**.

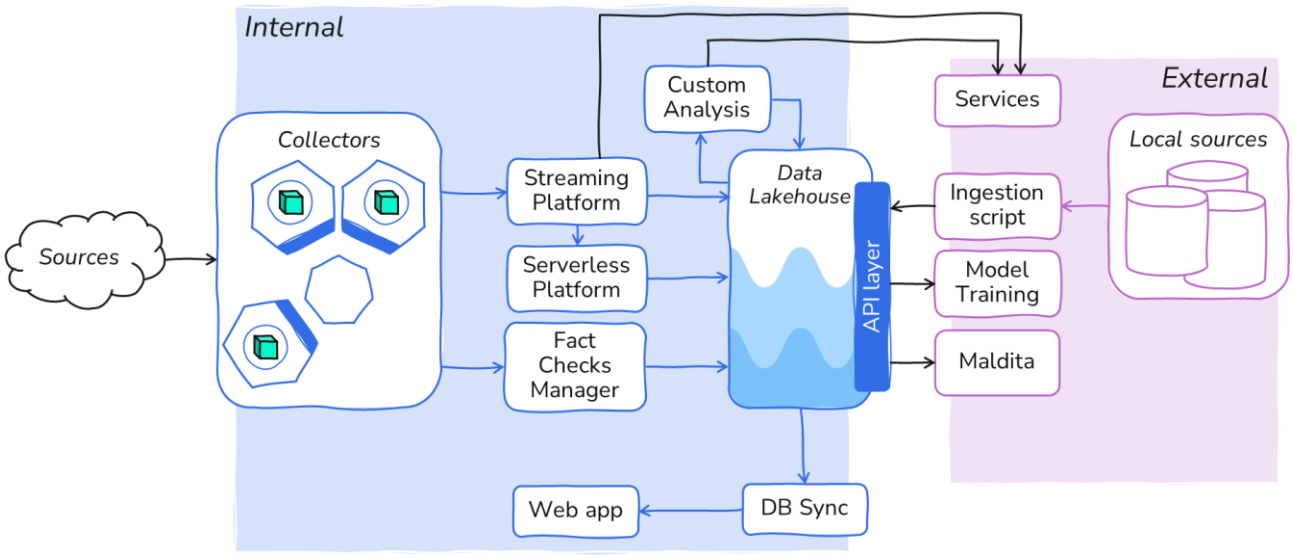

***Figure 1** – Architecture Overview*

As shown in Fig. 1, the platform will contain both internal (represented by the blue bounding boxes) and external components (represented by the purple bounding boxes).

The **internal components** will be based on a containerised cloud infrastructure, which will allow for seamless integration and updates of individual components without affecting the entire AI4TRUST platform. By adopting containerization (see req. #10 and #11), it will be possible to manage and maintain each component dynamically, generating a platform that can scale and adapt to evolving requirements and advancements in technology. **External components**, on the other hand, will be deployed on partners' premises, and their specifications will be introduced in the next sections.

By making this distinction, the project acknowledges the need to manage different needs (see req. #21 and #22), such as integrating already available services, safeguarding IP protection, and satisfying computational constraints (particularly regarding GPU usage), etc.

The internal data flow will be based on a set of **collectors** used to retrieve near-real-time data from various external **sources** (see req. #01 and #02). The data collection will be related to different topics, and done

according to a predefined set of keywords (see req. #03) in different EU languages (see req. #05). Each collector will subsequently ensure the proper routing of the collected data to the appropriate components of the platform, for further analysis.

Once the data are collected, it will be routed based on its type: it will be directed either to the **fact checks manager**, which will harmonise the fact checks data's structure and content before storage (see req. #07), or to a **streaming platform** for data preprocessing and standardisation (see req. #08 and #25). This platform will make use of AI tools and algorithms, utilising both **external services** deployed in partners' premises (see req. #20) and an internal **serverless platform** for data processing (see req. #16).

Subsequently, the collected data will be stored in a **data lakehouse**[4] to support large volumes of data (see req. #24), which could then be further analysed offline through a series of **custom analysis** iterations (see req. #18).

The **DB sync** will synchronise and standardise a subset of the data in the data lakehouse (see req. #25), which will be of interest to the users (i.e., media journalists and policy makers) and that will be shown to them through a **web application** (see req. #27 and #31). The web application will retain the synchronised data in its low-latency storage (see req. #28), which will enable its OpenAPI standardised backend (see req. #12) to properly serve its frontend.

Access to the data stored in the data lakehouse for both **AI model training** (see req. #19, #29 and #31) and for interaction with external platforms (see req. #06) will be regulated through an **API layer** (see req. #09), which will adopt an OpenAPI formalisation (see req. #12), limit the egress of RAW data (see req. #23) and provide access control mechanisms, ensuring compliance with the legal, ethical, and security standards (see req. #15). One of the external platforms that will make use of this API layer for accessing the AI4TRUST data is **Maldita**'s one[5].

External **local sources** will be ingested into the AI4TRUST platform using ad-hoc **ingestion scripts** (see req. #17 and #26), which will be executed on the partners' premises.

A more in-depth overview of the proposed architecture is provided in the next sections:

- Data Ingestion (sec. 3.2) describes how data will be collected by the **collectors** and how the external **local sources** will be ingested.
- Elaboration and Analysis (sec. 3.3) describes the **streaming platform** (sec. 3.3.1), the **serverless platform** (sec. 3.3.2), the **custom analysis** (sec. 3.3.3) and the **external services** (sec. 3.3.4).
- Data Storage (sec. 3.4) describes the **data lakehouse**, its technologies and its data structures.
- Data Synchronisation (sec. 3.5) describes the **DB sync** component.
- Web Application (sec. 3.6) describes the **web application** and its database, its backend and its user interface to allow the user to interact with the relevant information.
- API Layer (sec. 3.7) describes the **API layer** that will allow external access to the data lakehouse.

[4] The data lakehouse augments the traditional flat data lake with modern and powerful capabilities for managing data schemas, table management, transaction processing and data versioning, all in a highly integrated environment.
[5] Maldita.es is an independent journalistic platform by Maldita, project partner, specialised in fact checking and data journalism techniques to combat disinformation. (official website: Maldita.es)

3.2. Data Ingestion

The AI4TRUST platform will actively **monitor multimodal content** encompassing text, audio, and visual elements in **multiple languages**, and from **different sources**. Each data input will be inserted into the AI4TRUST platform based on its respective type, as illustrated in **Figure 2**.

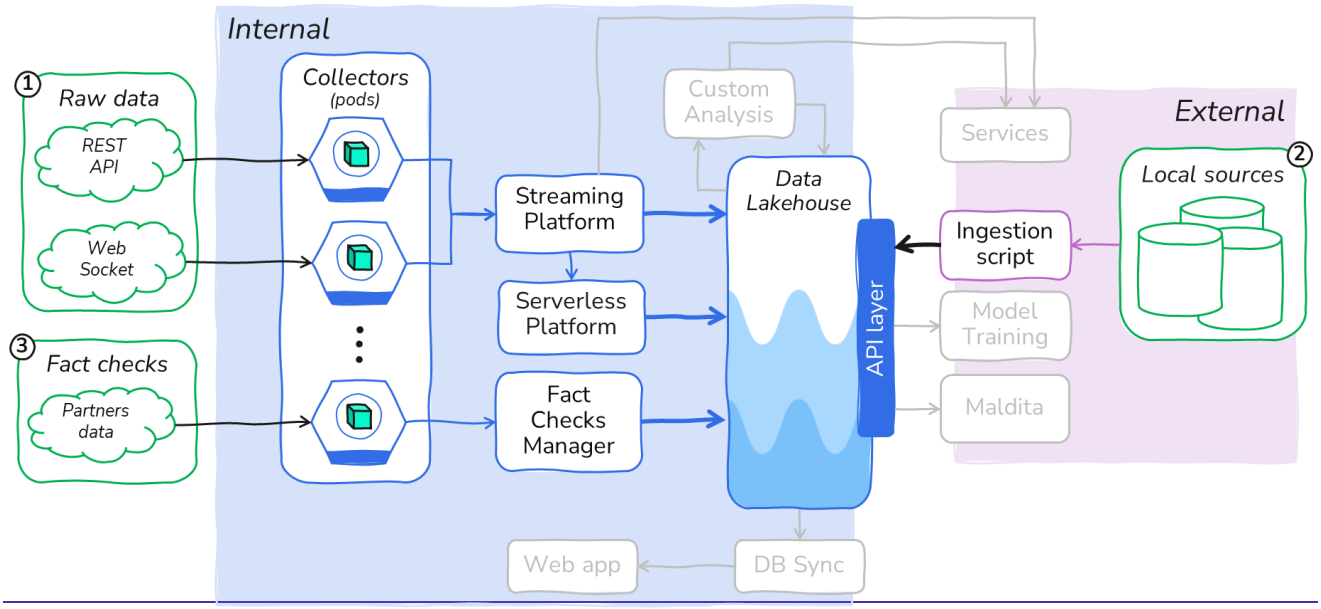

***Figure 2** – Data Ingestion*

The AI4TRUST platform will manage **three distinct types of input sources**, namely:

1. **Raw data sources:** original/non-processed data gathered from social media platforms, partners' platforms and web news feeds automatically collected through multiple scheduled collection processes.
2. **Local sources:** data gathered through a manually-based ingestion script, which have not been collected automatically before (e.g., data corpora collected in previous projects, externally processed data)
3. **Fact checks:** annotated/labelled data gathered from fact checkers' platforms through multiple collectors.

The **first type of input** defined as **raw data** consists of data sourced in near-real time from **online platforms** (web news feeds and news aggregators), **social networks** and **raw data residing on partners' data platforms** (e.g., Maldita's one). As of the current moment, our access includes Facebook via the Crowdtangle service and YouTube through their YouTube Researcher programme. Regrettably, the Twitter public API is presently insufficient in providing the necessary data volumes for meaningful scientific research. Consequently, we are contemplating the potential substitution of Twitter within our project with TikTok that is planning to open in the next few months its research API to European institutions. A second contingency plan is being considered, which also includes the purchase of a so-called 'basic' access right to the Twitter API.

Data from each source will be obtained through collectors (i.e., dedicated components for data collection), which use scheduled collection processes in case of a REST API approach, or web socket connections, in case

of continuous listening of data. The collectors will automatically gather the data according to the specified **topics and keywords** defined by the AI4TRUST consortium partners in the languages covered by our project. Each collected data item will then be transmitted to a **streaming platform for pre-processing and data harmonisation**, which will offer two possible processing paths: (i) send the data directly to the data lakehouse or (ii) route it to a serverless platform for further analysis. The collection and ingestion of data will be privacy compliant and done securely, according to the ethical, privacy, data protection and security requirements outlined in the AI4TRUST Data Management Plan (reported in D1.2).

A second **type of input** includes **manual input of unprocessed data, processed data or fact checks from local sources**, such as existing collections, experimental preliminary results and fact checking exports. This second data input path will be manually introduced into the AI4TRUST platform through ad-hoc ingestion scripts that may happen only once (in **Figure 2**, represented in the External section), as opposed to the collectors' approach which happens in a recurring fashion, thus enabling for customised inclusion of structured or unstructured data not yet available within the platform.

The **third and final data type of input** consists of **fact checks**, which represent the outcomes of fact checkers work. These fact checks will be acquired automatically, as opposed to the ones obtained from external local sources. To facilitate the integration and standardisation of fact checks data into the data lakehouse, a fact checks manager will be employed to homogenise the data, considering the different structures and standards.

3.3. Elaboration and Analysis

The AI4TRUST platform will envisage an **elaboration and analysis step** to process the ingested data before storing them in the data lakehouse. The ethical and legal framework, which will regulate the usage of the ingested data by the technology providing partners of the AI4TRUST consortium, was at this early stage described at a general level in Section 4 of this document and deliverable D1.2 - Data Management Plan, and will subsequently be detailed during the project. Furthermore, in compliance with the GDPR, “data controllers” (i.e., partners responsible for data processing compliance) and “data processors” (i.e., partners allowed to process personal data) will be appointed following the signing of the Data Protection Agreements attached to the above-mentioned D1.2[6].

To support the harmonisation, pre-processing and processing in nearly real-time of high volumes of data, the elaboration and analysis phase will be split across several processing steps:

[6] Art 26 GDPR requires differentiation between processing and joint control. In a controller-processor relationship, only the processor can process personal data based on documented instructions from the controller and must inform him of any relevant changes regarding processing. In most cases, commissioned data processing proceeds based on Art 28(3) GDPR which outlines minimum requirements, including the type of personal data to be processed, the object and purpose of processing, record keeping for audit (Art30 GDPR and, recital 82;); codes of conduct (art 40 GDPR); certification (Art 42 GDPR), general principles for transfers (art 44 GDPR) transfer on the basis of an adequacy decision (Art 45 GDPR), transfers subject to appropriate safeguards (art 46 GDPR), binding corporate rules (Art 47 GDPR); and right to compensation and liability (Art 82 GDPR). The controller is the first point of contact for the data subject and is responsible for data processing compliance but under Art 82 GDPR the processor is jointly liable with the controller.

- **Streaming Platform:** a platform that will support the nearly real-time preprocessing and harmonisation of the ingested data.
- **Serverless Platform:** a platform that will support nearly real-time analysis that might occur after the ingestion or the preprocessing process.
- **Custom Analysis:** a customised analysis (involving specific components of the overall AI4TRUST platform) that might occur at a later stage (i.e., not real-time).
- **External Services:** services that need to be executed on an external platform due to their hardware or IP requirements, and that will be accessed by the AI4TRUST platform through their APIs.

The relative details are explained in the following subsections.

3.3.1. Streaming Platform

The **Streaming Platform** is the heart of the (pre)processing layer, a key component which supports all the operations and data flows defined in the design of the AI4TRUST platform, handling all transformations and loading of datasets and models into the data lakehouse.

At the core, the streaming platform will receive from the ingestion layer all the data parcels, with the raw data embedded in the streaming in a properly defined way. The processing layer will then apply a Directed Acyclic Graph (DAG) of operations, which will be executed via atomic processors connected over the streaming topics in a mesh architecture. This approach will ensure **high performance, easy scalability and dynamic reconfiguration**, all core requirements for the platform.

Processors are written as serverless functions, which will receive a block of data in input (such as a single data entity or a group) and output the transformed data as a result. The design should follow the functional pattern, where processors do not keep state or conditions which could alter the processing, but are invoked in a stateless manner following the functional principles. Given a processor, by feeding a given input it should always produce the same output. This principle will ensure **proper reproducibility of all the data transformations**, a key aspect for data lineage and for supporting the re-processing of samples with updated processors, models or different DAGs.

Given that the data flow is managed via streaming topics, the processors will be executed in a dynamic function mesh, where their inputs and outputs will be bound to data streams by the platform itself, in a declarative manner. Furthermore, the functional design will enable the platform in delivering horizontal scalability, dynamically, based on the load on the various topics.

The **main components** are:

- Apache Pulsar[7], as the streaming platform;
- Pulsar Function mesh[8];
- Keda[9], as the load-based scaler;

[7] https://pulsar.apache.org/
[8] https://functionmesh.io/
[9] https://keda.sh/

- Kubernetes[10], as a computer platform.

At the end of every flow, a pre-configured **sink processor** will collect data associated to the predefined topics and properly persist the content both in the storage layer and in the platform catalogue.

Figure 3 presents an example graph depicting an ingestion and transformation pipeline for YouTube videos' comments:

- Raw data is ingested from the external YouTube API to collect videos' comments and pushed over the streaming platform.
- Comments are then processed in parallel, inside the mesh, via various functions to analyse the textual content, for example for hate speech detection, and also to download the thumbnail as an artefact.
- Raw data, thumbnails and the hate speech detector results are then united and properly stored in the data lake.

All the operations are performed inside containers, deployed in the mesh, and optimised for the specific task, while routing, data streaming and data persistence are handled transparently by the platform. While the image represents every step as a single block, the real deployment could scale every processor as needed to fulfil the required data rate, independently from all the other participants in the data flow.

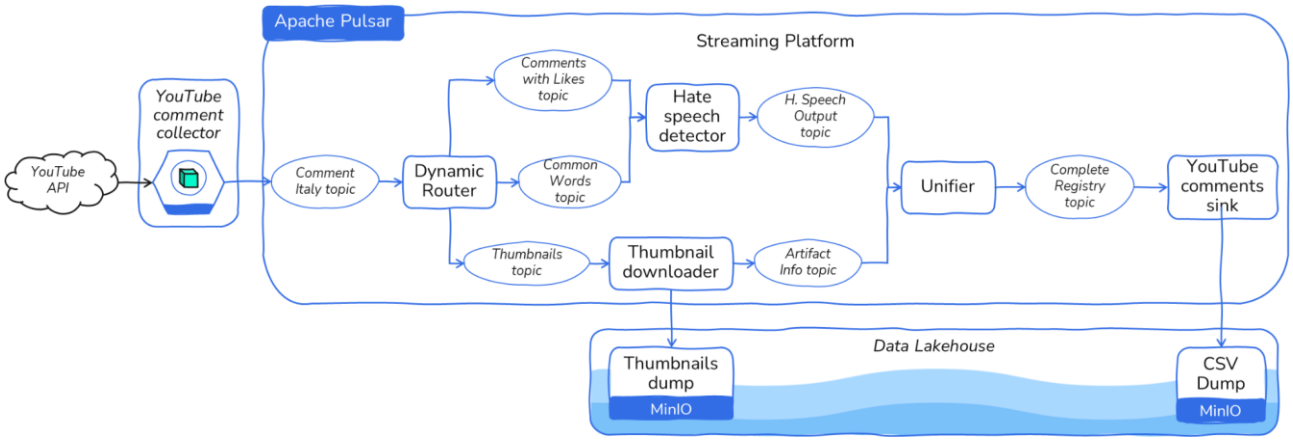

Figure 3 *– Example Streaming Pipeline*

3.3.2. Serverless Platform

The AI4TRUST platform leverages a powerful **Serverless Platform** to provide developers with the ability to write custom processors for executing dedicated logic during the data collection and processing flow. Developers can implement their custom logic according to the specific source, data and scenarios, focusing on the business logic and not on the operational side, while the serverless platform will handle the lifecycle of deployable functions, fulfilling all the operational capabilities such as deployment management, scaling, routing and networking.

The **core components** are:

[10] https://kubernetes.io/

- Kubernetes[11], as the compute platform;
- Nuclio[12], as the serverless platform.

In the AI4TRUST platform, the serverless layer is responsible for **running data processor**s, which apply transformations or elaborations on data. Given the highly specialised needs of every data source, the AI4TRUST platform design is aimed at enabling developers in writing serverless functions with minimal overhead, which will be fully managed by the serverless platform, from building a runnable image to deploying the component, setting up the environment, wiring the connections and collecting logs and metrics.

This approach aims at focusing developers and researchers efforts towards properly solving the specific problems, while all the boilerplate and management aspects should be handled by the platform itself, in a transparent and observable manner. This approach should lower the barrier to entrance for new developers, facilitate the growth of the AI4TRUST platform and also **lower the maintenance costs**.

AI models developed and used in the AI4TRUST project will be a **combination of current state-of-the-art implementations and future research and development**. As such, the AI4TRUST platform will need a strong integration with the widely adopted and used **Python AI** stack and the common **ML platforms**.

By adopting a serverless platform as the execution environment, an easy-to-use platform which can support a wide range of tools and libraries is obtained, all while ensuring that every AI processor is developed, deployed and used in a fully independent and isolated environment, with direct integration with the AI4TRUST platform.

A given AI model can be deployed, from source code, as a connected processor in the serverless platform, where it will receive input data samples and output the result on the return path, without requiring direct access and support for the streaming system or the data repository, effectively detaching the data access and the message routing from the actual code. This will reduce the code needed, facilitating the reuse of processors and ensuring **flexibility and rapid development**.

3.3.3. Custom Analysis

The **Custom Analysis** will be distinguished from the Streaming Platform and the Serverless Platform due to the fact that it might occur at a later stage (i.e., not real-time). In fact, custom analyses may be useful to execute **further processing of available data**, which may e.g., extend the results obtained in the streaming platform or may help in gathering additional relevant information useful to end-users.

3.3.4. External Services

The **External Services** refer to all the services that cannot be executed in the AI4TRUST platform, due to their specific hardware (e.g., GPU) or IP requirements.

[11] https://kubernetes.io/
[12] https://nuclio.io/

The AI4TRUST platform will use the **partners' External Services APIs** to access the respective functionalities and store the analysis results on the data lakehouse. Due to the fact that the requests to the External Services will be made from the AI4TRUST platform itself, no interaction with the API layer is foreseen.

In order to reduce the egress bandwidth from the AI4TRUST platform, the RAW data will not be sent by default to the External Services. This way only the **metadata** will be sent and it will be up to the External Services to "rehydrate" the data (i.e., obtaining the RAW data directly from the original source). If "rehydration" will not be possible, a policy to provide a limited set of the stored RAW data to the External Services will be defined.

As an example, the STT (speech-to-text) external service API will allow the AI4TRUST platform to ask for an audio transcription in one of the supported languages. In the requests, there will be a required field (e.g., asset id, token, etc.) used to identify the raw audio file to be fetched back from the original source by the STT service (i.e., file rehydration). If the raw audio file cannot be retrieved (e.g., it is not accessible, it has been deleted, etc.), the STT API will allow the platform to upload the internally stored copy of the raw audio file and obtain the results.

3.4. Data Storage

The **Data Lakehouse** is the repository for all kinds of datasets managed by the AI4TRUST platform. Following proper design principles, it does not impose a predefined, rigid schema on data formats or structures. Nevertheless, the data lakehouse fully manages schema-based datasets, with advanced capabilities, such as schema verification and evolution, consistent views and even point-in-time time travel capabilities, all while supporting transactions where needed. It is an open, scalable and dynamic object storage system, paired with modern data and table formats which ensure consistency, performance, flexibility and ease of use.

The **core components** are:

- Minio[13], as the open-source object store;
- Apache Parquet[14] and Apache Iceberg[15], as data and table format.

The base layer supports the **integration of various modern query engines**, even deployed together: the data consistency is ensured by the core layer and the access protocols used by query engines, such as Apache Arrow Flight[16]. The choice of query engines depends on specific needs and can be possibly fine-tuned and even dynamically managed on a per-user/per-scenario basis.

[13] https://min.io/
[14] https://parquet.apache.org/
[15] https://iceberg.apache.org/
[16] https://arrow.apache.org/

The platform will support Dremio[17], as a user self-service query and data product platform, but additional query engines such as Trino[18] (formerly PrestoSql[19]) or Apache Spark[20] could be easily added and connected to the same sources, delivering additional capabilities to the data lakehouse.

Given the domains analysed for the project, and the data sources currently proposed, it should be possible to define a **standardised structure for data sets**, which will help in properly managing both collected and generated data in a harmonised and compliant way. In order to properly model the data entities that will live inside the platform, it is needed to analyse the context of the project and conduct a detailed survey of the various datasets collected and features defined by AI models, with the objective of defining the proper standardisation procedure for every kind of content.

Nevertheless, it can be defined as a **base set of data models** that will be used for the project, by reducing all the various complex data types to a composition of basic data models centred around the content type:

- Core data;
- Base data: Text/Image/Audio/Video.

By defining a core model with identification, typing and tracking metadata, content-specific base models can be built for the various primitive data types: text, audio, video, image.

On top of this, **complex data types can be modelled** as a composition of base data models and core data plus additional properties, which can be modelled as explicit types following a schema or as unstructured key/value pairs. This will enable defining the processing and ML blocks which handle core/base data models, ensuring the reusability and composability of the processing stack.

Internal access to all the datasets stored inside the data lakehouse will be possible via the following interfaces:

- Query execution via the query engine;
- Direct file access for both artefacts[21] and data files;
- Custom REST read-only API access for specific data models, deployed on a per case approach to facilitate integration with external processors and partners.

Every stakeholder of the consortium will be able to access one or more interfaces, with proper management of credentials, authentication and authorization, which includes proper respect of data licences, usage rights and privacy concerns.

[17] https://www.dremio.com/
[18] https://trino.io/
[19] https://prestodb.io/
[20] https://spark.apache.org/
[21] Differently from the datafiles, which contains the actual data, the artefacts are used to store an auxiliary information, produced with the data processing workflows, such as logs, metrics, analysis reports, supporting intermediate datasets, etc.

3.5. Data Synchronisation

The **Database (DB) Sync. Component** (see **Figure 4**) will be responsible for synchronising a subset of the data lakehouse in a smaller database. In fact, the data lakehouse is optimised for storing large volumes of data at the cost of slower access, while low-latency access to a subset of the data shall be provided. For these reasons, the synchronisation process will be limited to the data that needs to be exposed in the UI. The DB sync. component will also be responsible for the data harmonisation, by defining standardised schemas for data storage, to ensure a uniform and consistent approach for handling data within the AI4TRUST platform.

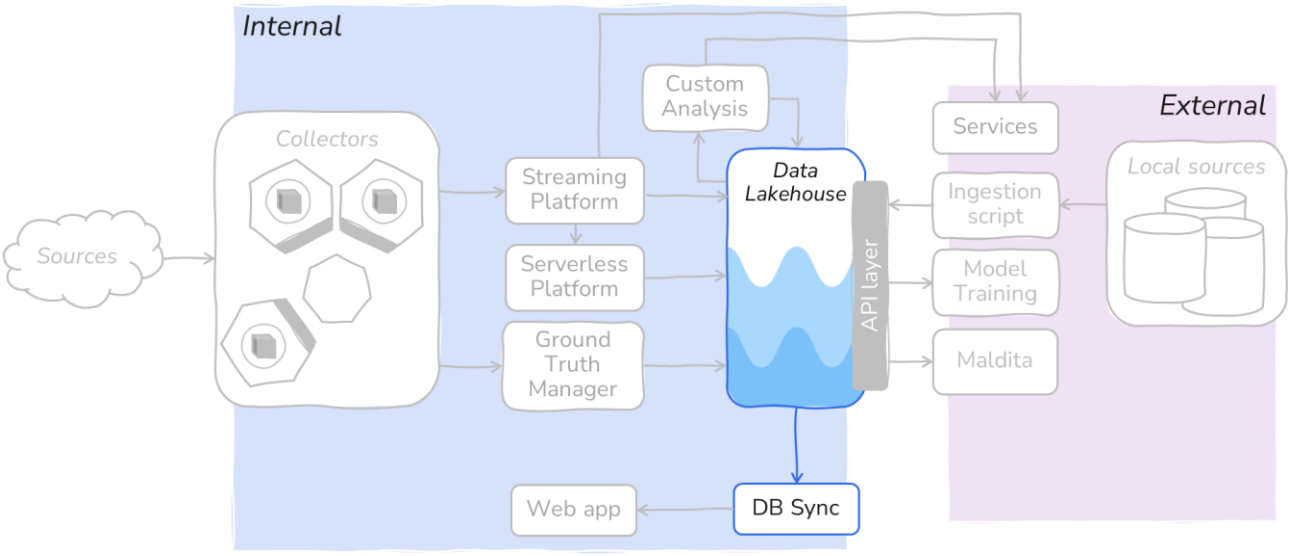

Figure 4 *– Database sync. Component*

3.6. Web Application

User interaction with data of interest will be made possible through the use of a structured **Web Application**. An in-depth visualisation of the main components of the web application is shown in **Figure 5**.

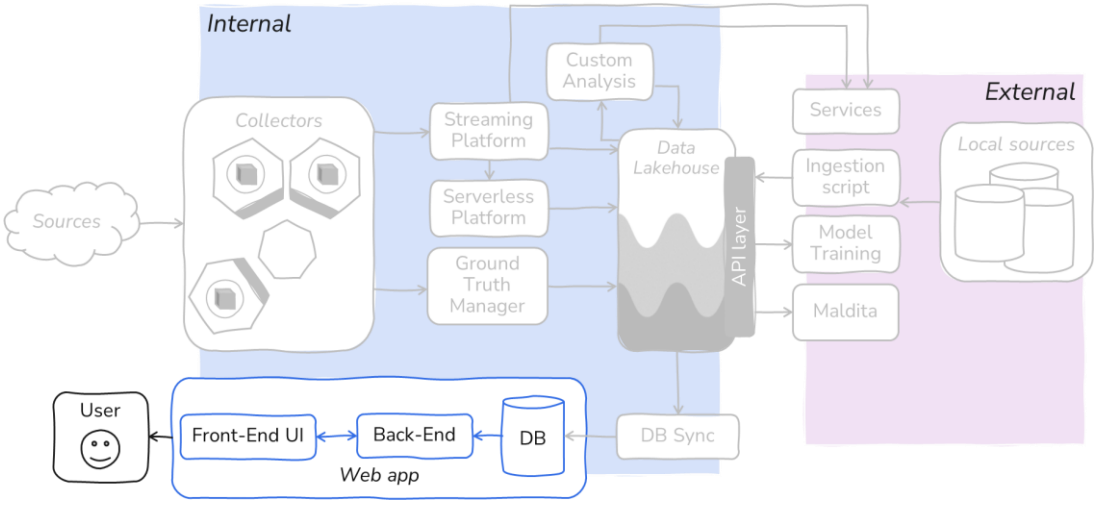

Figure 5 *– Web Application*

The **main components** will be:

- Database;
- Back-end;
- Front-end User Interface (Front-End UI).

To avoid preliminary calculations and intermediate results that may not be of interest to the final user, the **database** will contain only a subset of the analysed data stored in the data lakehouse. Furthermore, the use of a database for the web application will be necessary to provide low-latency access to the data of interest, avoiding long access times concerning the data lakehouse, due to its nature and the amount of data. As for the data that will be stored in this database, standardised data schemas will be defined to ensure consistency and facilitate seamless handling. These schemas will establish a consistent structure that governs the organisation and representation of the data within the AI4TRUST platform.

The **back-end** component will contain the implementation of APIs and the integration of necessary logic to support the **Front-End UI**. This implementation will enable users to interact with the data ensuring a seamless and user-friendly experience.

In the initial version, the **relevant data** exposed to the user will be **pseudonymised** and ideally **anonymised**, avoiding individual identification. At each iteration of the platform (see Section 5), the legal, ethical and privacy implications will be reviewed on an ongoing basis within the relevant WP teams in conjunction with SAHER Europe, which is AI4TRUST's partner responsible for legal and ethical issues.

3.7. API Layer

The **API Layer** will serve as a fundamental component for **external access to the data lakehouse**. Its primary objective will be to establish an authentication layer that restricts external access, controlling the number and type of data requests. The API layer will also act as an intermediary between external third parties and the internal S3-like API, which will enable researchers and consortium partners to access the datasets for any kind of need, all while respecting the proper usage and sharing policies associated with the content. Furthermore, the API layer will act as an intermediary between external third parties and the internal RPC-style API, enabling queries which will then be executed by the query engine on the actual data, and receiving the results in response. This approach will drastically reduce the need to share full datasets with partners, ensuring both control of infrastructure costs and the enforcement of access and usage policies on a per-user/per-organisation model. Finally, the API layer will enable a higher level of abstraction of data access and aggregation.

4. Ethics, Security, and Privacy Implications

Points #13 and #15 of Table 2 in Section 2 outline the AI4TRUST requirements in developing a proper methodology to ensure **legal, security and ethical compliance**, along with the **explainability of the AI developed**. The legal, ethical and security implications of this project extend to all aspects of data collection, processing, storage, internal and potentially external access control and credentials, as well as data destruction and/or re-use. **Section 3.3** provides a more comprehensive exploration of deployment issues in this regard. Throughout the project, maintaining ethical and legal compliance, along with ensuring strong security, will be of utmost importance. To accomplish this, engineers and the ethics-legal team will actively engage in continuous and proactive discussions, initiated from the project's outset and sustained throughout its whole duration.

The importance of ensuring the **explainability of the integrated AI solutions** will be addressed in these discussions recognising the changing regulatory landscape and the real-world concerns of civil society. The **ethical perspective** indeed highlights the importance of incorporating EU values and considers the subsequent implications and **societal impact** of any proposed solution. It is recognised that society is heterogeneous and in terms of the development of methodologies, this diversity is reflected in the challenges posed by multiple languages for fact checkers working on identical issues concerning the problem of classification of online disinformation and misinformation. The act of categorising online information presents indeed challenges that stem from the **diverse cultural interpretations of terms** that cannot be easily translated “word for word”. These challenges highlight the need to establish a shared understanding of the specific meaning, usage, and related synonyms of these terms. In order to address this, it is crucial to consider the significant legal, security and legal matters that arise from European Union (EU) legislation and regulation. It is advised to refer to **Deliverable 1.2 - Data Management Plan** alongside this document to fully grasp these key issues.

The AI4TRUST project then sheds light on concerns that affect **public trust in the digital world**, as seen through the lens of civil society. The primary concern for civil society is to discern **reliable technologies** and ascertain their **reliability, authenticity, credibility, and neutrality**. Moreover, key security, ethical and privacy implications are not sufficiently captured by existing legislative and regulatory frameworks but are embedded in a **fluid and constantly evolving regulatory landscape**, which the project is mindful of. They are, however, expressed in legal requirements (most notably the GDPR) regarding the use of personal data, with or without the subject’s explicit consent; data storage; data destruction; data corruption, onward sale and re-use (either in full or in part), purpose specification, purpose limitation and the potential impact on the personal autonomy, dignity, privacy and integrity of the individual whose data is being processed.

AI4TRUST will ensure that the handling of all data sets created, processed, or re-used will be informed by guidelines for **FAIR data management**. Consequently, personal data processing will align with **Regulation (EU) 2016/679 (GDPR)**. Briefly, the GDPR is concerned with data that can directly or indirectly identify an individual. It clearly differentiates pseudonymised data (personal data) from anonymised data (not personal data because it undergoes irreversible changes to prevent identification).

AI4TRUST project specifically utilises **pseudonymised, anonymised, synthetic data** which does not rely on, require, or permit, the identification of individuals and/or their activities. Each partner is thus responsible for ensuring the application of the GDPR and regularly reviewing, assessing, and updating their legal and ethical compliance, overseen by SAHER Europe. Each partner is moreover responsible for following ethics, security and data protection at their premises.

Accordingly, the **AI4TRUST Data Management Plan (D1.2)** will determine the data points at which any transition from pseudonymised to anonymised data may be undertaken for both internal and external use. This complies with GDPR requirements.

The AI4TRUST Consortium has reached a consensus on implementing a range of technical and organisational measures, clearly defined in the **Consortium Agreement** (CA - Annex to D1.1 - Project Management Plan):

- Pseudonymise online data where anonymisation is considered impossible. Therefore, the following subsequent measures will be implemented: (i) perform the research, specifically by collecting and utilising online data in the most minimal and necessary manner (referred to as the 'Data Minimisation Principle'); (ii) secure that online data originates from public sources; (iii) ensure that all legal requirements are implemented, including requirements stemming from the GDPR and from national legislation. This entails establishing an effective legal basis for processing the data. Additionally, prior to sharing any personal data, it is necessary to execute appropriate data sharing agreements, such as Joint Controllership Agreements or Data Processing Agreements.
- Discuss other operational, managerial, and technical measures to be implemented.
- Monitor in a continuous manner the compliance in the use of personal data. This includes a careful assessment of all data sets, to assess the risk of re-identification and whether they can be deemed anonymised based on legislation and state-of-the-art literature.
- Destroy collected data within 12 months of the completion of the project to allow for the end of project dissemination and follow-up, unless specifically requested and agreed.

More specifically, the GDPR clearly prioritises the use of **pseudonymisation as a valuable tool** for data protection and management purposes. Pseudonymisation indeed renders data records unidentifiable, effectively reducing the risk of unauthorised access or exposure of sensitive information. However, it also allows authorised data processors and controllers to access and manage the data. The AI4TRUST project accords with the GDPR's recommendations on using pseudonymised data:

- Art 4(5) GDPR defines pseudonymisation as "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".
- Art 6(4) GDPR (Recital 26) permits pseudonymised data to be reversed with a proper key by authorised personnel. Unintentional pseudonymised data leakage may incur legal consequences for the host organisation and have harmful, adverse effects on the individual human subject. Therefore, guaranteeing the security of pseudonymised data must be a high priority for those using and/or sharing it for whatever purposes.
- Recital 28 GDPR: "The application of pseudonymization to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations."

In summary, the GDPR considers pseudonymisation as a suitable method to achieve the following objectives: (i) safeguard data subjects and data controllers while ensuring compliance; (ii) protect personal data on legacy systems against unauthorised access; (iii) keep a backup in order to temporary store original data when personal data is being anonymised. It ultimately signals data protection by design and default (as

required by Art. 25 GDPR). **Data subjects' right to be forgotten** is also seen to be enhanced by pseudonymisation, especially when automated. The processes for complying with this are to be addressed.

In addition, the GDPR expects data controllers and organisations to: (i) **protect individuals' data** and places responsibility on them for evaluating risks (under Recital 83, committing them to implement **mitigating measures, such as encryption**), (ii) c**ombat risks by ensuring a high level of security** that is suitable for the identified risks, as required by Article 32 of the GDPR to ultimately ensure **confidentiality, integrity and resilience processing**. In order to accomplish these objectives the AI4TRUST project will focus on addressing and continuously evaluating the requirements for extracting data both for and from fact-checking purposes, as well as from processing.

Ethical considerations and adherence to the values and principles of the European Union serve as the foundations for all EU funded projects. Consequently, the processing of data must adhere to the highest and rigorous ethical standards and the applicable EU, international and national law on ethical principles. Ongoing AI4TRUST ethical reflection expands upon legal compliance and involves the application of a series of standard questions as well as more specific questions related to its Work packages and Tasks. These relate to:

- The use of information and context specification;
- Privacy;
- Human autonomy and integrity (e.g., proportionality of AI apps);
- Control, influence, and power (e.g., transparency, accountability, responsibility);
- Impact on social contact patterns;
- Gender, minorities, and justice;
- Human values;
- Sustainability;
- Bias and viability.

Partners are thus required to **assess the potential risks associated with data processing activities**, particularly those involving large-scale processing of personal data or extensive monitoring of publicly accessible areas. AI4TRUST personal data processing will be specifically safeguarded by: (i) identifying the kind of personal data collected, the partners involved in the processing of personal data, as well as the legal basis declared by partners, (ii) implementing technical and organisational measures (iii) data processing agreements where needed.

Whereas the legal and regulatory frameworks on pseudonymisation and anonymisation are specific, **new challenges arise from the ethical and security perspectives** from the opportunities and potential for harm arising **from the deployment of AI**.

Data masking for instance is seen as a form of pseudonymisation but must be strict in terms of the fields exposed to ensure that the data cannot be associated with a specific individual. Critics note that "pseudonymisation may permit identification using indirect means. When a pseudonym is used, it may be feasible to identify the individual concerned by the data by analysing related data". This point holds great significance as it encompasses ethical requirements regarding the right of the individual human data subject to integrity, autonomy, dignity as well as privacy. It also raises issues regarding data minimisation, purpose specification and limitation and non-linkability and reuse for unknown and/or unspecified purposes, again intrinsic to ethical considerations.

The AI4TRUST platform employs a range of methodologies and technological measures in order to **guarantee adherence to the above-mentioned privacy and data protection principles**, outlined by the GDPR:

1. The **Data ingestion phase** (as described in **Section 3.2** and to be implemented in Task 2.2) will be conducted in a privacy-compliant and secure manner, aligning with the ethical requirements outlined in the AI4TRUST D1.2 - Data Management Plan. Indeed to ensure the safeguarding of user identification, pseudonymisation techniques, such as hashing, will be effectively employed;
2. In the **Data elaboration phase** (described in **Section 3.3**), the roles of "controllers" and "processors" will be specified in compliance with GDPR regulations. The "controllers" will be partners who are responsible for ensuring data processing compliance, while the "processors" will be partners authorised to process personal data.
3. **Data access** (as described in **Section 3.4**) will be managed through the implementation of appropriate authentication and authorization techniques. In order to achieve this, every stakeholder within the Consortium will be granted access to one or more interfaces. This access will be managed to ensure the proper handling of credentials, authentication, and authorisation. It is essential to emphasise that this management also encompasses a diligent adherence to data licences, usage rights, and privacy concerns.
4. **Data exposed** to end-users (as described in **Section 3.6**) will undergo pseudonymisation to prevent user identification, this will be implemented also thanks to the pseudonymization techniques adopted in the Data ingestion phase. Indeed the user identity information is actually determined by the data imported during the Data ingestion process. If the data is already pseudonymized at the source, there is no need to implement additional techniques when generating the output.

Ultimately the AI4TRUST project lifecycle aligns with and closely corresponds to **significant legislative and regulatory initiatives undertaken within the EU, including the forthcoming AI Act**. Special consideration will be given to related and pertinent legislation, as well as the positions and recommendations made by the EDPS and other relevant civil society organisations.

5. Conclusions and Next Steps

This deliverable provides a comprehensive **architectural overview of the AI4TRUST platform**, detailing its main components and their interactions. It also explored the storage and manipulation of data, delving into the critical considerations surrounding privacy, ethics, and security issues. The ultimate objective is to ensure that the AI4TRUST platform fulfils ethical, privacy, and data protection obligations and requirements, including AI explainability (as defined in WP1 and WP4). More in-depth implementation details and specifics will be covered in the forthcoming deliverables **D5.5 - AI4TRUST Platform v1** due on M18 (June 2024), **D5.6 - AI4TRUST Platform v2** due on M32 (August 2025), and **D5.7 - Final AI4TRUST Platform** due on M38 (February 2026).

References

- Art. 29 Data Protection Working Party (2017) Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679
- Art. 29 Data Protection Working Party (2014), Opinion 05/2014 on Anonymisation Techniques. Available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
- ENISA (2019) Pseudonymisation Techniques and Best Practices, https://www.enisa.europa.eu/news/enisa-news/enisa-proposes-best-practices-and-techniques-for-pseudonymisation, https://enisa.europa.eu/topics/cybersecurity-policy/data-protection
- European Commission (2022) The 2022 Code of Practice on Disinformation https://digital-strategy-ec.europa.eu/en/policies/code-practice-disinformation
- EPRS (2021) Regulating facial recognition in the EU, Brussels. https://europarl.europa.eu/RegData/etudes/IDAN/2021/698021/EPRS_IDA(2021)698021_EN.pdf
- European Union, Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02 https://www.refworld.org/docid/3ae6b3b70.html
- EDRI (2022) Position Paper Respecting Fundamental Rights in the cross border investigation of serious crimes. https://edri.org/wp-content/uploads/2022/10/EDRi-position-paper-Respecting-fundamental-rights-in-the-cross-border-investigation-of-serious-crimes-7-September-2022.pdf
- European Group on Ethics in Science and New Technologies (2018) An Ethical, societal, and fundamental rights dimension for the EU policies, Brussels
- European Data Protection Board (2022) Guidelines on Art.60 (GDPR), Guidelines on dark patterns in social media platforms interfaces, toolbox on essential data protection safeguards for enforcement cooperation between EEA and third country SAs, https://edpb.europa.eu/news/2022/edpb-adopts-guidelines-art-60-gdpr-guidelines-dark-patterns-social-media-platform-en
- General Data Protection Regulation (GDPR) regulation (EU)2016/679 General data Protection regulation OJ L 119,04.05.2016, in force as of 25.05.2018, https://gdpr-info.eu
- Iapp. (2019). Publicly available data under the GDPR: Main considerations. [online] Available at: https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations/
- Independent High-Level Expert Group On Artificial Intelligence (AI HLEG) (2020), The Assessment List For Trustworthy Intelligence, Artificial Intelligence (ALTAI) For Self-Assessment, 17 July 2020. Doi:10.2759/002360. See also: https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines
- Privacy International (2023) Artificial Intelligence https://privacyinternational.org/learn/artificial-intelligence