WellFort: A Platform for Privacy-Preserving Data Analysis

by Tomasz Miksa, Tanja Šarčević, Rudolf Mayer (SBA Research) and Laura Waltersdorfer (Vienna University of Technology)

Data has become deeply ingrained in all phases and aspects of industrial and scientific research. The potential for new discoveries based on data-driven research is growing fast, due to the high volume and granularity of personal data collected by individuals, e.g., by means of ubiquitous sensors and IoT devices. However, small and medium-sized organisations typically face challenges in acquiring and storing personal data, particularly in sensitive data categories.

To enable organisations to leverage the full potential of the personal data they collect, two main technical challenges need to be addressed: (i) organisations must preserve the privacy of individual users and honour their consent, while (ii) being able to provide data and algorithmic governance, e.g., in the form of audit trails, to increase trust in the result and support reproducibility of the data analysis tasks performed on the collected data.

Organisations could further improve their analysis by integrating data from different, complementary sources and users, but there is no established way to request such data – and individuals often refrain from sharing data due to lack of trust. We believe that individuals would contribute their data to research and welcome services that enhance their experience, for example, in the fitness or medical domain, if security and privacy were guaranteed and users could maintain control by giving explicit consent.

On the other hand, organisations must be able to provide evidence that the data is used according to the consent given, and to collect information on how the analysis was performed. This information must encompass the traditional provenance, such as who and when data was accessed, but also information on software libraries and scripts used to analyse the data. This is especially important in litigation cases and scientific peer review when new claims are scrupulously evaluated. Privacy-preservation cannot be the reason for not making the data analysis auditable.

To address these problems, we are developing a platform called WellFort, which provides secure storage for users' sensitive data, while delivering a trusted analysis environment for executing data analytics processes in a controlled privacy-preserving environment. A novelty of our approach is that organisations do not have direct access to data, but only allow this in aggregated or anonymised form. Organisations can benefit from a large group of individuals that are potentially willing to share their data for research. Users benefit from a privacy-preserving and secure platform for their data, and can contribute to research projects in a secure manner. Finally, scientific researchers have a detailed source of microdata, if data subjects give consent to their research proposals.

The conceptual architecture of the platform is depicted in Figure 1. There are three distinct actors:

- Users store their data in the platform, give consent to analyse it, etc. They use an application provided by the organisation and interact with the

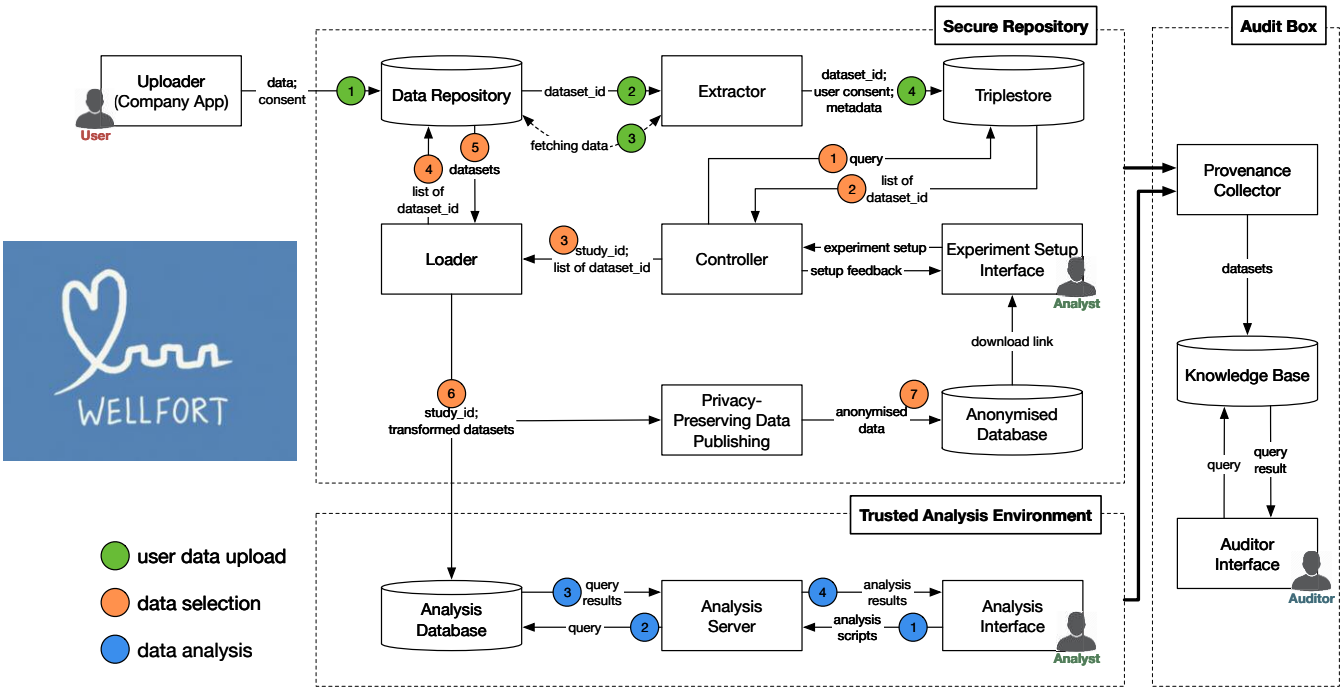

Figure 1: The WellFort architecture, comprising the Secure Repository, Trusted Analysis Environment and the Audit Component.

platform using a dedicated user interface.

- Analysts can run experiments on the platform. They define which types of data will be used and perform the actual analysis.
- Auditors can analyse evidence collected to answer specific audit questions that depend on the purpose of the audit, e.g., a litigation case. A special form of auditor is a user wanting to know when and by whom their data was used.

The architecture consists of three component groups (each marked with dashed lines in Figure 1), each serving a different purpose:

- Secure Repository – stores data uploaded by a user, together with a fine-grain consent [1], and allows the selection of data to be used in experiments by the analyst.
- Trusted Analysis Environment – selected data that fulfils experiment criteria, e.g., consent, fit for purpose, etc. is duplicated to this component for further analysis. This component provides mechanisms to conduct data analysis in a privacy-preserving manner, e.g. using DataShield [2]. Data selection is usually expressed via queries.
- Audit box – collects and manages provenance data to support auditability [3]; it can be accessed to answer audit-related questions on personal data access and usage.

Figure 1 further depicts three processes that may be executed in the platform:

- User data upload – starts when a user's application sends data to the platform. It extracts metadata from the data, and stores it together with the data and the consent indicated during the upload in the platform. Thus, every dataset uploaded to the platform is linked to a minimal set of information that allows for its retrieval.
- Data selection – analysts define the search criteria for data they want to use in their experiments. If the platform has enough data fulfilling their criteria (and consent for usage), then the process loads the actual data into the Trusted Analysis Environment. Analysts do not have access to individual datasets. The search for data relies only on high-level information provided in the metadata.
- Data analysis – analysts process the data and produce results by submitting code to the platform. The platform will ensure that the analysts will not be able to identify or infer data subjects from the analysis.

The platform is currently evaluated in the medical domain, with two start-ups providing and analysing medical and wellbeing data. These two data sources are analysed individually, and are also integrated to provide a more holistic view on patients. In the future, we plan to focus on use cases in other domains, to evaluate whether our approach extends well to other types of data and analysis processes.

The WellFort project receives funding from the Austrian Research Promotion Agency FFG under grant 871267 (WellFort).

Link:

[L1] https://kwz.me/h6X

References:

[1] J.D. Fernández et al.: "User consent modeling for ensuring transparency and compliance in smart cities", Personal and Ubiquitous Computing (2020), 1–22.

[2] A. Gaye, Y. Marcon, J. Isaeva, et al.: "DataSHIELD: taking the analysis to the data, not the data to the analysis." International journal of epidemiology 43.6 (2014): 1929-1944.

[3] R. Mayer, T. Miksa, and A. Rauber: "Ontologies for describing the context of scientific experiment processes"; in Proc. of the 10th Int. Conf. on e-Science, Guarujá, SP, Brazil, 2014.

Please contact:

Tomasz Miksa
SBA Research, Austria
tmiksa@sba-research.org