The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2690-6090.htm

Received 14 December 2021
Revised 10 December 2022
8 February 2023
Accepted 15 February 2023

A privacy-preserving federated learning architecture implementing data ownership and portability on edge end-points

Patience Mpofu
Department of Computer Engineering, University of Zimbabwe, Harare, Zimbabwe
Solomon Hopewell Kembo
Centre for ICT Product Development Training and Services, Harare, Zimbabwe
Marlvern Chimbwanda
Department of Computer Engineering, University of Zimbabwe, Harare, Zimbabwe
Saulo Jacques
Department of Research and Development, Hacking Ecology, Barcelona, Spain
Nevil Chitiyo
Department of Research and Development, Clamore Solar, Harare, Zimbabwe, and
Kudakwashe Zvarevashe
Department of Analytics and Informatics, University of Zimbabwe, Harare, Zimbabwe

Abstract

Purpose – In response to food supply constraints resulting from coronavirus disease 2019 (COVID-19) restrictions, in the year 2020, the project developed automated household Aquaponics units to guarantee food self-sufficiency. However, the automated aquaponics solution did not fully comply with data privacy and portability best practices to protect the data of household owners. The purpose of this study is to develop a data privacy and portability layer on top of the previously developed automated Aquaponics units.

Design/methodology/approach – Design Science Research (DSR) is the research method implemented in this study.

Findings – General Data Protection and Privacy Regulations (GDPR)-inspired principles empowering data subjects including data minimisation, purpose limitation, storage limitation as well as integrity and confidentiality can be implemented in a federated learning (FL) architecture using Pinecone Matrix home servers and edge devices.

Research limitations/implications – The literature reviewed for this study demonstrates that the GDPR right to data portability can have a positive impact on data protection by giving individuals more control over their own data. This is achieved by allowing data subjects to obtain their personal information from a data controller in a format that makes it simple to reuse it in another context and to transmit this information freely to any other data controller of their choice. Data portability is not strictly governed or enforced by data protection laws in the developing world, such as Zimbabwe's Data Protection Act of 2021.

Practical implications – Privacy requirements can be implemented in end-point technology such as smartphones, microcontrollers and single board computer clusters enabling data subjects to be incentivised

International Journal of Industrial Engineering and Operations Management
Vol. 5 No. 2, 2023
pp. 118-134
Emerald Publishing Limited
e-ISSN: 2690-6104
p-ISSN: 2690-6090
DOI 10.1108/IJIEOM-02-2023-0020

© Patience Mpofu, Solomon Hopewell Kembo, Marlvern Chimbwanda, Saulo Jacques, Nevil Chitiyo and Kudakwashe Zvarevashe. Published in *International Journal of Industrial Engineering and Operations Management*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and no commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

whilst unlocking the value of their own data in the process fostering competition among data controllers and processors.

Originality/value – The use of end-to-end encryption with Matrix Pinecone on edge endpoints and fog servers, as well as the practical implementation of data portability, are currently not adequately covered in the literature. The study acts as a springboard for a future conversation on the topic.

Keywords Data privacy, Data portability, Federated learning, Peer-to-peer networking

Paper type Research paper

1. Introduction

Traditionally, machine learning (ML) involves training data in public data centres. In addition to latency and power consumption concerns, centralized ML presents privacy challenges (Brecko *et al.*, 2022). Federated learning (FL) is a paradigm shift that brings public programs to private data as opposed to bringing data to the programs. Developing a large corpus of distributed data for Internet of things (IoT) edge and fog computing use cases requires supporting FL infrastructure. Our previous work developed an "offline-first" architecture for household aquaponics (Mpofu *et al.*, 2021). In this paper, we extend the previously developed "offline-first" architecture into a peer-to-peer (P2P) setup consisting of a Bluetooth-powered overlay network based on Matrix protocol. The fully distributed architecture is developed on Matrix Pinecone which utilizes an overlay Bluetooth network to relay ML updates on low-cost Android smartphones. The architecture enables data portability for data owners. Data protection legislatures in the Global South, such as Zimbabwe's Data Protection Act of 2021, do not provide rigorous guidance and enforcement of data portability (Data Protection Act, 2021). Consequently, the study utilized Article 20 of the General Data Protection and Privacy Regulations (GDPR) to evaluate the proposed architecture. GDPR principles, including data minimisation, purpose limitation, storage limitation as well as integrity and confidentiality, were implemented in the FL architecture.

1.1 Data protection

The IoT continues to evolve with an incredible impact on industries and human life, driven by smart devices, smartphones, cloud computing and intelligent applications. IoT is made up of millions of clients exchanging massive amounts of critical data; subsequently, as the data generated grows, so do risks of privacy invasion and breaches. Privacy risks are compounded when data are controlled by centralized entities that include cloud service providers and social networks. Consequently, several physical, regulatory, legal and technical remedies to privacy breaches have been proposed, tested and implemented.

It is against the backdrop of increased privacy threats and breaches that Zimbabwe formally enacted the Data Protection Act (DPA) in December of 2021. Central to DPA is a thrust on data privacy and protection for data collected by handlers within and outside the country. However, the act does not adequately address the issue of data portability. Alternative privacy enforcement guidelines to complement DPA to strengthen data owners' need to own and port their own data, therefore, becomes necessary. One of the most rigorous and widely cited data protection frameworks is the GDPR.

With the development of technology and the phenomenal growth of Internet-based malpractice, the European Union (EU) realized the necessity of new safeguards. In 1995, the EU passed the European Data Protection Directive, which established baseline criteria for data privacy and security and served as the foundation for implementing laws in each of the member states. However, the Internet was already changing and becoming the data hoover it is today. During the last four decades, Internet-based innovations attracted data portability issues that necessitated legislative changes. The first Internet banner advertisement emerged in 1994. Most financial institutions provided Internet banking in 2000. Facebook had its public debut in

2006. A Google user sued the corporation in 2011 because it had scanned her emails. Two months later, the EU's data protection body declared that the 1995 regulation needed to be updated and called for "a comprehensive approach to personal data protection" (Wolford, 2020).

To ensure more uniform consumer and personal data protection across EU member states, GDPR standards are applicable to all EU members. The GDPR's main privacy and data protection regulations include the following: requesting subjects' permission before processing their data, using anonymization to safeguard acquired data's privacy, notifying users of data breaches, managing the cross-border flow of data in a secure manner, mandating the appointment of a data protection officer to manage GDPR compliance for certain firms. Simply defined, the GDPR requires a minimum set of requirements for businesses that handle the personal data of EU citizens to better protect the processing and movement of that data (Wolford, 2020). In addition, Article 20 (1) of GDPR specifically empowers individuals or data subjects to have control over their personal data as enshrined in the 8 fundamental rights accorded to data subjects. The GDPR will be used as a best practice that the Global South countries like Zimbabwe should emulate.

This work was necessitated by the weaknesses of deploying ML models on the cloud where privacy is a concern. When storing personal information on the cloud, users put it at risk because they do not know exactly what the providers can do with the data. Personal information should be made inaccessible to users without the right authority to access it. A key security risk that prevents many people from using cloud services is the lack of trust between users and cloud service providers or cloud database service providers regarding the data. Data loss is possible, and it is possible for the user's data to be kept at a geographical location outside of the legal jurisdiction, which causes the user to worry about local law enforcement's legal accessibility and the rules regarding data stored outside of their territory. In light of this, it is crucial to safeguard user personal data utilizing decentralized ML, where information is secured and encrypted (Lord, 2014).

The goal of this study is to design a federated learning architecture utilizing the FedML open platform, where a delegated algorithm is implemented, and to develop a privacy-preserving architecture employing edge end-points that utilize the Lorawan network and the Pinecone peer-to-peer network. This architecture's aim is to assess ownership and portability while comparing data privacy to the GDPR's right to portability.

1.2 Data privacy and ownership opportunities

Regulatory and legal mechanisms to enforce privacy such as GDPR and DPA emphasize that at the centre of data protection is the data subject and not the companies that control the data that include data controllers and processors (van der Sloot *et al.*, 2019). However, most data subjects will possibly not understand the abstract and technical language contained in regulatory and legal documents, thus missing out on opportunities availed by novel concepts such as data portability and ownership. Additionally, regulatory and legal recourse to privacy violations in the Global South is generally reactive in nature as its effects are most prominent when a privacy breach has occurred. This paper proposes that a more proactive approach that enables users to unlock the value of their data as well as protect it entails privacy at IoT and ML end-points in a decentralized architecture.

The prevailing downward trend in the costs of storage, computing, software and hardware opens up opportunities to actualize data portability in viable and realistic ways. While these have privacy and security concerns to this current study, it must be acknowledged that some of the benefits of data portability include:

(1) Increased consumer control: the ability to utilize and monetize data without assistance from third parties and free from vendor lock-in.

(2) Unlocking value: by making the most use of the democratization of data, data subjects will own and deploy their data to diverse data applications as opposed to value restricted by being bound to a single set of data controllers or processors.

(3) Fosters competition and innovation: when data controllers and operators realize that data subjects own and port their data, operators and processors will be compelled to be competitive through innovative propositions and incentives to data subjects.

1.3 Objectives

The specific objectives of the study include:

(1) To study prominent data protection regulations, specifically DPA in Zimbabwe and GDPR in the EU, with a specific emphasis on data portability opportunities arising from their enactment.

(2) To develop an IoT architecture that utilizes decentralized IoT and ML in order to realize the opportunities availed by data portability.

(3) To assess the developed FL architecture against rigorous data portability regulatory requirements with the GDPR rules as a baseline for evaluation.

Following this introduction, the remainder of the essay is structured as follows: Section 2 reviews the literature on data protection and data portability laws in Africa and the GDPR of the EU. Additionally, it discusses the drawbacks of centralized ML, emphasizes the benefits of decentralized ML and proposes the idea of a peer-to-peer matrix. The third section, known as methodology, is where the research science research is put into practice to provide knowledge that experts in the relevant field may utilize to create solutions to their practical difficulties. A detailed proof of concept is presented. To wrap up our work, the results are examined in part 4, and the conclusion is presented in section 5.

2. Literature review

In this section, we review past and current work relating to data protection regulation and data portability rights, privacy challenges of traditional ML and privacy-preserving peer-to-peer networking.

2.1 Data protection regulation and data portability state of art

African countries have made efforts to enact legal frameworks due to the increased use of personal data, for data protection and data governance. However, the African Union does not have central legislation that governs the concept of data and safeguards its member countries. This is unlike the EU which enacted the GDPR. There is no single comprehensive data protection framework in Africa (Babalola, 2022). Where there are frameworks, there is either a lack of institutional enforcement or the enforcement in place is inadequate.

Some African frameworks are discussed below. Namely, African Union Convention on Cyber Security and Personal Data Protection 2014 (Malabo Convention), the Supplementary Act on Personal Data Protection within the ECOWAS (ECOWAS ACT) and the Southern African Development Community (SADC) Model Law on data protection.

(1) African Union Convention on Cyber Security and Personal Data Protection 2014 (Malabo Convention). It is Africa's first international data protection treaty, which became law in 2014. The Convention seeks to, among other objectives, harmonize member states' data protection laws and encourage them to develop frameworks to protect personal data across the continent. Unfortunately, the Convention's sufficiency

may be called into question because it omitted the data protection authority; definition of key concepts such as pseudonymization and cross-border processing, as well as the right to file a complaint with the regulator, data portability, data subject restriction on future processes. This omission could lead to cause conceptual confusion, particularly in cross-border enforcement situations (Babalola, 2022).

(2) Supplementary Act on Personal Data Protection within the ECOWAS (ECOWAS ACT). The Economic Community of West African States (ECOWAS) was formed for promotion of regional cooperation among member states, particularly for economic growth. The Supplementary Act A/SA.1/01/10 on the protection of personal data within ECOWAS (the Act) was enacted in 2010 to govern data protection within the member countries (ECOWAS, 2010). Just like the Malabo Convention, the Act also omits key terms such as processing and pseudonymization, personal data breach, cross-border transfer, complaint rights with the regulator, the right to data portability and so on.

(3) Southern African Development Community (SADC) Model Law on data protection was implemented due to the growing need for a standardized set of information policies in Sub-Saharan African countries. The Act was pushed by African, Caribbean and Pacific countries into law in 2013 (Titre du rapport, 2018). Although the law requires data transfer to take place only between SADC members or non-members with adequate data mechanisms, it does not provide SADC members or nonmembers with adequate data mechanisms.

The Zimbabwean Data Protection Act requires all data controllers to take appropriate technical and organizational measures that are necessary to protect data from unauthorized destruction, negligent loss, unauthorized alteration or access, and any other unauthorized processing of the data. It does not mention the data portability which might cause conceptual confusion.

The challenge of traditional ML is that machines are connected to the Internet when learning and that it plays a significant role in the development of ML systems. The fact that the data are online makes it prone to hackers which may trick the ML machines' algorithms into accepting bogus inputs by progressively conditioning them to give incorrect outputs (Lohn, 2020). Additionally, since the cloud is hosted online, the ML platform can allow hackers to deceive the system with malicious inputs and trick computers into thinking something is real when it is not by presenting them with false information (Gupta *et al.*, 2022). Such attacks can have devastating effects because they have the potential to be both long-lasting ML security risks.

2.1.1 Insufficient training data. The lack of both quality and quantity of data is a big challenge when employing ML algorithms and can lead to a security threat of predicting the wrong information. Despite the fact that data are critical in the processing of ML algorithms, many data scientists believe that insufficient data, noisy data and unclean data are particularly taxing on the algorithms. The following elements can have an impact on data quality and potentially risk the correct predictions; Noisy data are to blame for erroneous predictions, which have an impact on decision-making and classification accuracy. Incorrect data are also to blame for erroneous programming and EU's GDPR's primary goal is to improve individuals' control and rights over their personal data, as well as to simplify the regulatory environment for international business. The regulation contains provisions and requirements related to the processing of personal data of individuals and applies to any enterprise – regardless of its location and the data subject (Wolford, 2020). Unlike other data protection regulations from Africa, it includes data portability which is a legal right (Article 20 of the GDPR). It states that, in some cases, a user may be able to obtain personal data from a data controller in a format that allows the user to reuse your information in another context, and to freely transmit this data to another data controller of the user's choice. This is known as the right to data portability (Data Protection Commission, 2023).

2.2 Privacy challenges of traditional machine learning

Traditionally, ML for the IoT was accomplished by uploading all data from each connected device to the central cloud to train a generic model that could be distributed and used on all devices. The following security issues arise when using central servers in traditional ML:

2.2.1 Data privacy. In ML, it is essential to preserve the privacy and confidentiality of massive datasets sent to the central server when the data are already included in the ML model itself. This type of ML makes it possible for attackers to carry out covert data extraction assaults in this circumstance by putting the entire ML system at risk. To defend themselves, organizations must implement policies that attempt to prevent function extraction attacks while also protecting ML systems from data attacks.

2.2.2 System manipulation online. Another ML model outcome. As a result, erroneous data may have an impact on the accuracy of the results.

The aforementioned security threats prompted the development of federated architecture, which aims to address the above issues.

2.3 Federated learning

FL is a new ML paradigm that has emerged as a result of data privacy concerns. It was pioneered by Google in 2017 (Google, 2017) as an alternative to the centralized, standard ML approach. FL is realized by obtaining high-quality centralized ML models from training data that is distributed over a large number of clients utilizing even the most unreliable and slow Internet connections (Konecny *et al.*, 2017). The FL workflow can be summarized as having three distinct processes that start with decentralized training of data by clients. Clients then upload the model to a server for aggregation. The final step involves the server sending model updates back to clients.

FL has also partly been inspired by new regulatory requirements such as the GDPR which aims at giving users more control over their personal data (Cheng *et al.*, 2021). Consequently, individuals or clients will only need to upload model updates but not raw data to a central server where the models are aggregated. With the use of secure aggregation protocols, the update parameters will not leak user information to the server (Cheng *et al.*, 2021).

In order for the federated systems to be effective, two factors that include heterogeneity and autonomy need to be prioritized when setting up the FL systems. Heterogeneity refers to the differences existing among participating clients in terms of data, privacy, requirements and tasks. Autonomy property within FL systems allows organizations or individuals to retain control by being able to decide whether or not to associate or disassociate themselves from FL systems. Robust FL systems must be able to tolerate entry and departure into the system (Li and Wen, 2019). Li and Wen (2019) propose a categorization of FL systems into data partition, model, privacy level and communication architecture. FL is a relatively new concept that is part of a larger attempt to increase privacy while also giving ML models access to fresh data that would otherwise be unavailable. FL is a ML system in which a number of dispersed nodes use their locally stored data to develop a shared prediction model collectively. Because training data is not transferred to a central server, it can give better data privacy. Federated learning is particularly suited for edge computing applications, as it can take advantage of edge servers' CPU capacity as well as data acquired from widely dispersed edge devices. It is necessary to overcome a number of technical hurdles in order to create such an edge-federated learning system.

The local computing element of federated learning lends itself well to edge and fog computing paradigms. Consequently, a number of FL deployments have utilized edge and fog clusters for local storage and computation. Hussain *et al.* (2019) proposed the deployment of edge computing in oil rigs to process latency-intolerant tasks. The edge computing components would become part of a federation and provisioned from nearby or mobile micro

data centres to augment the edge units' limited capacities. Such a solution is robust as it captures uncertainties in computation and communication within the federated environment. Another project that also leverages edge computing (Javed *et al.*, 2020) developed IoTEF architecture made up of an Apache-Kafka pub/sub platform and Kubernetes fault-tolerant management platform which was implemented in a smart buildings project. In developing the next-generation innovation platform for healthcare-related organizations (Long *et al.*, 2021), collaborated while keeping sensitive health data private and also complying with regulatory obligations that include the Health Insurance Portability and Accountability Act of 1996 (HIPAA).

PERFED-CKT (Cho and Wang, 2021), is a personalized FL framework that caters to clients using heterogeneous model architectures and does not directly send their parameters. PERFED-CKT managed to produce high model performance without incurring huge communication costs. **Ψ**-*Net* is a federated learning framework that established a firm structure-information to re-align models that have been chaotically fused as a result of heterogeneous sources. The heterogeneity of different privacy settings has the potential of producing better model performance whilst not compromising users' privacy (Li and Wen, 2019).

Li and Wen (2019) identifies different types of autonomy necessary for developing a robust FL system. Association autonomy is the ability of a participant member to associate or disassociate itself from FL systems. As such an FL system's robustness should be gauged by its tolerance to entries into and departures from the FL system. Another important form of autonomy defined is communication autonomy. Communication autonomy is the ability to determine how much information is communicated to others. Participants should have the ability to handle dynamic size communication during the communication process. Members within the system should be in a position to deal with the "Privacy Paradox" which entails gaining more value from sharing more information which in turn increases risks of the private data being violated.

There exists two major ways of communication in FL systems: centralized design and decentralized design (Li and Wen, 2019). In the centralized setup, one server is responsible for the aggregation of data from other participants and updating parameters to the global model. The centralized design whilst easy to implement poses a lot of privacy risks and unfairness. A decentralized design, whilst challenging to implement, is ideal as it treats all participants equally. Model-centric and data-centric federated ML are the two forms of federated ML. Let us start with model-centric because it is more frequent right now.

2.3.1 Model-centric federated learning. 'Model-centric' refers to any FL solution to deliver better centrally administered models. Data are generated on a local level and is kept decentralized. Each client keeps track of its own information and is unable to access the information of other clients. Data are not distributed in a uniform or independent manner. The most likely case for data-centric is when a person or organization has data in PyGrid that they want to safeguard (instead of hosting the model, they host data). This would enable a data scientist who is not the data owner to seek training or inference on the data (OpenMined/PyGrid, 2021). In the model-centric world, there is usually a pre-configured, pre-trained model ready to be modified, especially when horizontal data is involved. ETL, analysis, experimentation and model selection are all processes in the ML pipeline that have already been completed.

2.3.2 Data-centric federated learning. However, under the data-centric approach (Gooday, 2020), methodologies and tools will be required to allow sufficient data discovery, wrangling and preparation in order to fully harness the power of the data in the network. At the same time, any data leakage must be measured, limited and controlled. Differential privacy and PSI are examples of concepts and techniques that can help, but automating control to satisfy the needs of the various parties involved, as well as assuring data governance and regulation compliance, is no minor feat.

2.4 Privacy preserving peer-to-peer networking: matrix protocol

The federated ML on a Matrix protocol is an open standard for decentralized, real-time communication over IP that is interoperable. There is an open standard in the shape of the Matrix Specification. It is interoperable, which means it is designed to work with other communication systems, and it is an Open Standard, which means it is simple to figure out how to work with it. Matrix is decentralized, which means there is no central point – anyone may host their own server and have complete control over their data. It is meant to work in real-time, making it suitable for developing systems that demand immediate data interchange, such as instant messaging. Each user is connected to a single server, which they refer to as their "home server". Because each Matrix server federates with other Matrix servers, users can join in rooms created on any Matrix server. This implies that you can communicate with anyone on any server. You can also host your own server, providing you complete control over your data. Self-hosting also allows you to customize your server to meet your specific needs, such as bridging to other chat networks (such as IRC, XMPP, Discord, Telegram, and so on) or hosting bots.

Synapse (Matrix Foundation, 2021c), is the most stable and widely used implementation of Matrix. Developed in Python, it has a huge memory footprint and is therefore not suitable for development on small devices.

Dendrite (Matrix Foundation, 2021a), written in Golang, is a second-generation Matrix home server. It is a more reliable, scalable and efficient alternative to Synapse. It also has a very small footprint. It is targeted for home server deployments as well as P2P in-browser nodes and mobile phones. It can be executed in polylith or in the recommended monolith modes.

Yggdrasil (Matrix Foundation, 2021d) is an end-to-end encrypted IPv6 network that uses a name-independent routing scheme. It is truly P2P as it works in an entirely ad-hoc manner with no points of centralization. It also possesses self-healing and scalable attributes.

Pinecone (Matrix Foundation, 2021b) is an experimental overlay routing protocol suite designed for global end-to-end connectivity over different types of media including TCP, WebSockets and Bluetooth Low Energy in a multi-hop peer-to-peer manner.

2.5 Contributions

We present a proof of concept that implements a data portability-supporting architecture utilizing Matrix Pinecone home servers. A distributed overlay network provides the networking infrastructure for the underlying federated learning application suitable for the IoT-based urban farming use case.

2.6 Summary and synthesis

Whilst the GDPR provides a solid framework to enforce data privacy, it relies on the regulator's ability to monitor and penalize data processors and data controllers. Using a proactive approach in which data subjects have control of implementing data ownership and portability on end devices, we identify GDPR principles that are within the control of data subjects out of the seven proffered in the framework. Table 1 below illustrates the analysis of the GDPR principles and the extent to which the data subject has control.

It is apparent from the analysis that four principles, namely purpose limitation, data minimization, storage limitation, and integrity and confidentiality, are within the control of the data subject. The analysis guides the development of an architecture that is implementable at the end-point allowing for data subjects to own their own data enabling them to port the data at will. End-point components suitable for implementing this data subject-oriented approach include edge computing, federated learning and peer-to-peer network infrastructure.

3. Methodology

In this research, the methodology employed is design science research (DSR). The DSR is a paradigm for problem-solving that seeks to improve human understanding through the creation of physical products (Hevner, 2007). It is also a research methodology that generates information about both the object of the design process and the method utilized to create a product. The main goal of the DSR is to produce knowledge that professionals in the relevant industry can use to develop answers to their real-world problems. DSR is particularly noteworthy in the disciplines of engineering and computer science.

3.1 DSR as an artefact

An IT-enabled artefact, privacy-preserving FL architecture, is developed and evaluated in this article using the DSR approach. The methodology is helpful for solving issues in the real world where an iterative design process is used to create an IT-enabled artefact. The DSR mainly consists of three interconnected cycles: relevance, rigor and design. The DSR cycles are illustrated in the diagram.

3.2 Relevance cycle

Our earlier work, in which we built three automated households Aquaponics systems to guarantee food security, served as the foundation for the privacy-preserving FL architecture. For the automated aquaponics systems (AAS), the bulk of the data processing and storage in our earlier work was on an offline-first edge and fog computing. The edge and fog architecture protected the homeowners' data because it was processed and stored at community servers made possible by edge computing. The concept relied on nearby homes using a community mini-data centre to maintain data sovereignty and privacy.

The community was involved in the process of choosing and evaluating the most pertinent smart services that should be developed for AAS, and their ideas were captured at the relevance stage. In order to develop a perception, data from the environment was gathered using a bottom-up strategy. This approach is suitable for this research because it is incremental and has allowed us to expand and improve our earlier work 'Using an IoT Edge and Fog Architecture with Privacy Preserving in Automated Home Aquaponics' to this work and allowed us constant refinement of our project over time (Mpofu *et al.*, 2021).

3.3 Rigor cycle

This rigor cycle requires examining the knowledge base, which contains supporting theories, approaches, domain experiences and expertise. The privacy-preserving federated architecture presents different technologies such as ML, networking, software engineering, and hardware engineering, edge and fog computing. The common objective of all these distinct technologies is to produce an intelligent user experience. We based the design of the

Table 1. Analysis of GDPR principles

GDPR principles							
	Lawfulness fairness and transparency	Purpose limitation	Data minimisation	Accuracy	Storage limitation	Integrity and confidentiality	Accountability
Data Subject	No Control	Can Control	Can Control	No Control	Can Control	Can Control	No Control

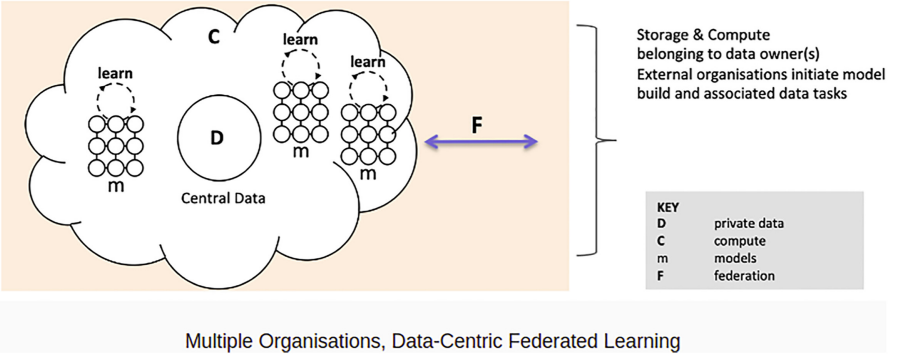
Source(s): Author's own; Table by author

privacy-preserving federated learning on knowledge from the field of expertise (electricians and farmers to assist in the creation of the AAS). The DSR's iterative nature helped our project to be successful due to its flexibility to build design artefacts allowing us to refer to the knowledge base while building the architecture (see Figure 1).

3.4 Design cycle

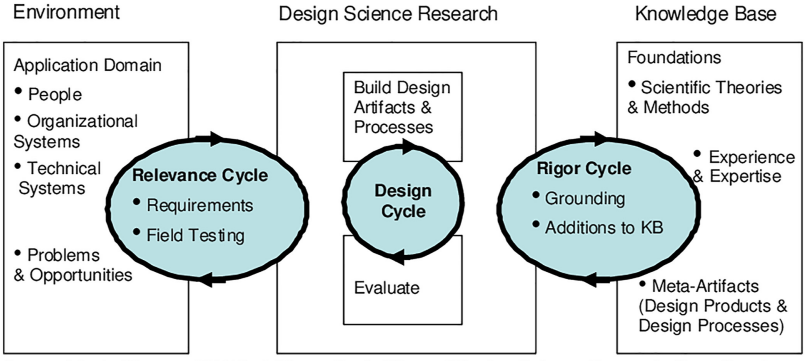
The research design comprised multiple design cycles conducted to test application functionality, improve the design and achieve the requirements from the knowledge base and the environment (Figure 2).

The design was completed in three cycles. We came up with a prototype of an architecture utilizing Arduino microcontrollers in the first design cycle. The architecture consisted of edge computing devices that detected water leaks in aquaponics systems as well as electroconductivity, pH and temperature sensors. During this cycle, a cluster of Raspberry Pi single-board computers was used to implement fog computing acting as a small local data centre. The lightweight Kubernetes distribution k3s, which controls the NodeRed and MongoDB Docker containers, is also hosted by the fog. To evaluate the system, the ten



Source(s): A. Gooday. (September 21, 2020). Understanding The Types Of Federated Learning. Open Mined. https://blog.openmined.org/

Figure 1.
Data-centric federated learning



Source(s): A. Hevner. (2007). A Three Cycle View of Design Science Research. Scandinavian Journal for Information Systems. https://aisel.aisnet.org/cgi/viewcontent.cgi?article = 1017&context = sjis

Figure 2.
DSR interconnected cycles: relevance, rigor and design

makerspace's technical staff tested the system. After a flaw in the architecture was discovered, we continued to improve it until the testers approved the systems.

The second cycle allowed the deployment of a delegated algorithm to the Federated ML open platform to place focus on the network infrastructure and edge layers made up of the Lorawan and Pinecone peer-to-peer network.

The Pinecone network is added to the preexisting layers in the final cycle. It is composed of inexpensive Android cell phones. In order to relay data updates till they reach the k3s raspberry pi cluster, Bluetooth is used by the smartphones running Matrix Pinecone in the background to locate other Bluetooth-enabled smartphones. The FedML platform then gathers the local updates to provide a better global update after receiving the updates. Android cellphones with Bluetooth capabilities transmit global updates to the Arduino microcontrollers.

3.4.1 Proof of concept. The proof of concept consists of three layers: the FL algorithm layer, the network infrastructure layer and the edge layer. Deployment of the algorithm was delegated to the FedML open platform in order to place focus on the network infrastructure and edge layers made up of the Lorawan and Pinecone peer-to-peer networks, respectively.

The edge devices in the architecture utilize the FedML AI platform (https://open.fedml.ai) for training and deployment of edge models with one-line commands. Algorithms for the fog layer utilize the private cloud deployment with Docker mode. The architecture utilizes on-device anomaly data detection for the federated IoT platform on FedML's App Ecosystem (https://open.fedml.ai/platform/AppDetails?id=137) to implement the sound anomaly detection of the water pumps. FedML allows for the development of the federated ML algorithm through a "one-click import" based on community results and uses the application directly without intensive development circles.

3.4.2 Network infrastructure layer. Figure 3 shows all the architectural components making up the design. In utilizing the "security by design" paradigm, the IoT Aquaponics system uses "offline-first" edge and fog computing for most of the data storage and processing. The edge and fog architecture preserves the household owners' data as it can be

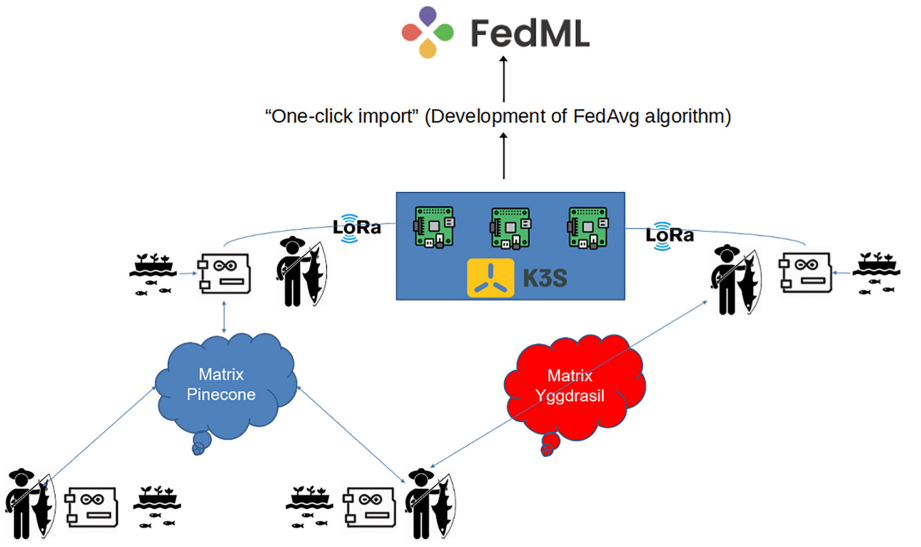


Figure 3.
System architecture: FL algorithm layer: FedML open platform

Source(s): Author's Own

stored and processed at the household level. Households within the community can opt to pool their data together and have it managed at a local mini-data centre as opposed to utilizing the cloud, in the process ensuring privacy and data portability.

There is a need to constantly monitor whether the water pump is running. Failure to detect abnormal water pump sounds may result in the vegetables not receiving water containing the fertilizing fish waste. We utilized a ML model that compares the normal sound of the water pump to the prevailing sound coming through. If the microcontroller detects a difference in sounds, which would be an anomaly, a message is sent to the system alerting the owner of the unit, otherwise, no action is taken. The offline-first approach ensures that an alert is only sent to the fog or cloud when a sound anomaly is detected. Consequently, data are securely retained within the edge network, in the process saving on bandwidth and preserving privacy.

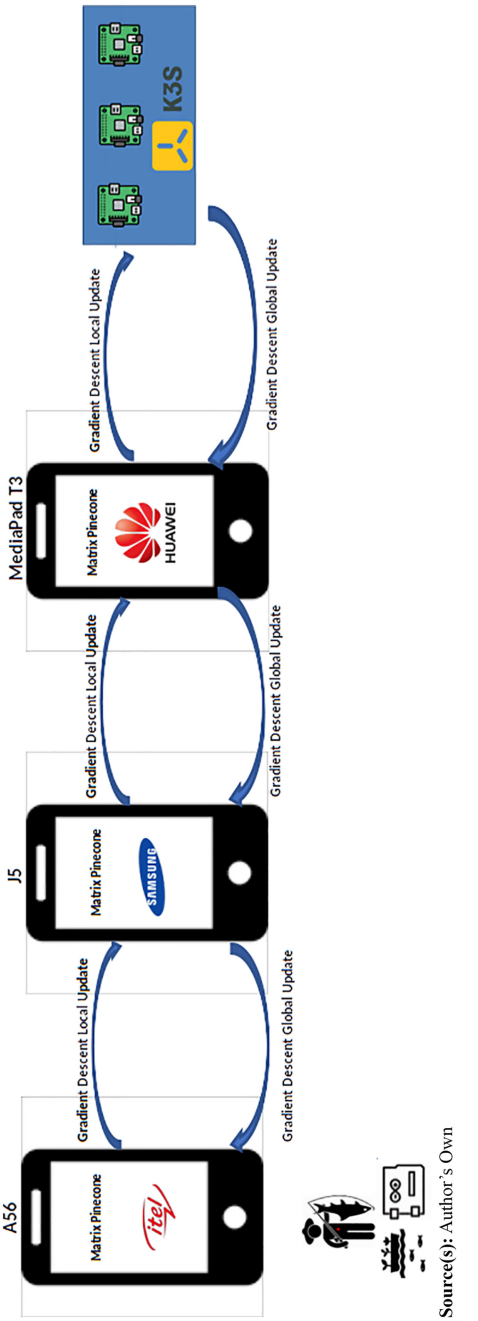
The Arduino microcontrollers, which provide edge computing services, have a water leak, electroconductivity, pH and temperature sensors attached to them. Fog computing is implemented as a cluster of Raspberry Pi single-board computers operating as a mini-data centre within the community. A LoRaWAN network glues together the household Aquaponics units creating the fog layer. The fog also hosts k3s, a lightweight Kubernetes distribution that manages the NodeRed and MongoDB docker containers.

3.4.3 Edge layer. Figure 4 shows a Pinecone network made up of low-cost Android smartphones. Using Bluetooth, the smartphones running Matrix Pinecone in the background identify other Bluetooth-enabled smartphones in order to relay data updates up until the data updates reach the k3s raspberry pi cluster. The updates are then forwarded to the FedML platform which aggregates the local updates developing an improved global update. The global updates are sent back to the Arduino microcontrollers through Bluetooth-enabled Android smartphones.

The iterative first-order optimization technique known as gradient descent is used to locate the local minimum or maximum of a given function (GD). The iterative process continues until the data are transmitted to the K3s cluster which forwards the local updates to the FedML open platform, where it is trained and returned to each node iteratively in its best possible state. The data are transferred from the k3s cluster to the local minimum, which is better than its adjacent nodes in our example to the Itel phone. TLS is used to encrypt all data delivered through Pinecone from beginning to end. Additionally, messages are cryptographically signed for validity. However, the protocol is still in its early stages, so there may be potential theoretical assaults that we are unaware of.

4. Results analysis

Table 2 illustrates an analysis of the different architectural components of the proposed system against GDPR principles enabling the implementation of data portability. The analysis demonstrates that all four key data portability supporting principles of GDPR are implemented in the proof of concept. Arduino microcontrollers implement data minimisation by only uploading local updates instead of all the data collected which are received by the FedML open platform. The limited local updates uploaded by microcontrollers also ensure that the FedML possesses virtually unusable global updates in the process ensuring purpose limitation. Storage limitation principle ensures that data are kept in a form that permits the identification of data subjects for no longer is necessary for purposes for which the personal data are processed. Uploading local updates also ensures storage limitations. Pinecone network encrypts data using Datagram Transport Layer Security (DTLS), and LoraWAN used at the k3s cluster uses AES encryption ensuring that the fourth principle of integrity and confidentiality is implemented.



Source(s): Author's Own

Figure 4.
Pinecone architecture

	Data portability principles			
	Data minimisation	Purpose limitation	Storage limitation	Integrity and confidentiality
Microcontrollers	Local updates	Local updates	Local updates	N/A
Pinecone Network	N/A	N/A	N/A	DTLS
K3s cluster	N/A	N/A	Global updates	AES Encryption
FedML	Global updates only	N/A	Global updates only	N/A

Source(s): Author's own

Table 2. Data portability principles

5. Conclusion

Previously developed IoT applications hosted on edge and fog infrastructure enabled a privacy-preserving alternative to cloud services. However, the design still required the cloud for ML applications that required multiple training datasets to develop ML models. Our presented FL architecture ensures that ML tasks requiring multiple training datasets can be executed without pushing all the data to the cloud. Autonomy and heterogeneity, which promote privacy and interoperability respectively, are key requirements for any FL application. Matrix infrastructure, which our study imposed on top of the FL application, natively has autonomy and heterogeneity features, thus strengthening FL applications in the architecture. The Matrix layer also makes privacy management fine-grained as it can be activated locally at the edge using the Pinecone implementation. GDPR has stood out as the most comprehensive data protection framework. The proposed FL architecture was evaluated against GDPR data portability-related principles of data minimisation, purpose limitation, storage limitation as well as integrity and confidentiality. The FL architecture implemented all four data portability requirements at the three layers of the architecture, namely the FL algorithm, infrastructure and edge layers.

The most prominent approach to data protection is the use of regulatory mechanisms targeting corporate players. Whilst penalties attached to the violation of privacy deter corporate entities from abusing data subjects' privacy, its reactive approach does not present opportunities for data subjects to derive value from their own data. The proactive approach demonstrated in this study which is proactive and technical and grounded on data portability ensures that data subjects can benefit from unlocking value from owned data whilst fostering healthy competition among data controllers and processors whilst preserving privacy.

References

Babalola, O. (2022), *Data Protection Legal Regime and Data Governance in Africa: An Overview*, Africa Portal, available at: https://www.africaportal.org/publications/data-protection-legal-regime-and-data-governance-africa-overview/

Brecko, A., Kajati, E., Koziorek, J. and Zolotova, I. (2022), "Federated learning for edge computing: a survey", *Applied Sciences*, Vol. 12 No. 18, doi: 10.3390/app12189124.

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T. and Papadopoulos, D. (2021), "SecureBoost: a lossless federated learning framework", *IEEE intelligent Systems*, Vol. 36 No. 6, pp. 87-98, doi: 10.1109/MIS.2021.3082561.

Cho, Y.J. and Wang, J. (2021), Personalized federated learning for heterogeneous clients with clustered knowledge transfer.

Data Protection Act (2021), "65384-T cyber & data protection Act.indd", *Postal and Telecommunications Regulatory Authority of Zimbabwe*, available at: https://www.potraz.gov.zw/wp-content/uploads/2022/02/Data-Protection-Act-5-of-2021.pdf

Data Protection Commission, D. P. C (2023), "The right to data portability", (Article 20 of the GDPR). Data Protection Commission, available at: http://www.dataprotection.ie/en/individuals/know-your-rights/right-data-portability-article-20-gdpr

ECOWAS (2010), "'Untitled.' Economic community of West African States (ECOWAS)", available at: https://ccdcoe.org/uploads/2019/10/ECOWAS-10216-Supplementary-Act-on-electronic-transaction.pdf (Accessed 9 December 2022).

Gooday, A. (2020), "Federated learning", OpenMined, September 21, available at: https://blog.openmined.org/federated-learning-types/

Google (2017), "Federated learning: collaborative machine learning without centralized training data", available at: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

Gupta, C., Johri, I., Srinivasan, K., Hu, Y., Qaisar, S.M. and Huang, K. (2022), "A systematic review on machine learning and deep learning models for electronic information security in mobile networks", *Sensors (Basel)*, Vol. 22 No. 5, p. 2017, doi: 10.3390/s22052017.

Hussain, R.F., Salehi, M.A., Kovalenko, A., Feng, Y., Semiari, O. and Fields, A.S.O. (2019), "Federated edge computing for disaster management in remote smart oil fields", *IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, Vol. 3, pp. 929-936, doi: 10.1109/HPCC/SmartCity/DSS.2019.00134.

Javed, A., Robert, J., Heljanko, K. and Främling, K. (2020), "IoTEF: a federated edge-cloud architecture for fault-tolerant IoT applications er", *Journal of Grid Computing*, Vol. 18 No. 15, pp. 57-80, doi: 10.1007/s10723-019-09498-8.

Konecny, J., McMahan, B., Yu, F., Suresh, A. and Bacon, D. (2017), Federated Learning: strategies for improving communication efficiency, pp. 1-10.

Li, Q. and Wen, Z. (2019), Federated learning systems: vision, hype, and reality for data privacy and protection, pp. 1-16.

Lohn, A.J. (2020), "Estimating the brittleness of AI: safety integrity levels and the need for testing out-of-distribution performance", ArXiv, abs/2009.00802, doi: 10.48550/arXiv.2009.00802.

Long, G., Shen, T., Tan, Y., Gerrard, L. and Clarke, A. (2021), Federated learning for privacy-preserving open innovation future on digital health.

Lord, N. (2014), "An expert guide to securing sensitive data: 34 experts reveal the biggest mistakes companies make with data security", *Digital Guardian*, available at: https://digitalguardian.com/blog/expert-guide-securing-sensitive-data-34-experts-reveal-biggest-mistakes-companies-make-data

Matrix Foundation (2021a), "Dendrite Matrix", available at: https://github.com/matrix-org/dendrite

Matrix Foundation (2021b), "Matrix Pinecone", available at: https://github.com/matrix-org/pinecone

Matrix Foundation (2021c), "Synapse Matrix", available at: https://github.com/matrix-org/synapse/

Matrix Foundation (2021d), "Yggdrasil Matrix", available at: https://yggdrasil-network.github.io/

Mpofu, P., Kembo, S., Jacques, S.M. and Chitiyo, N. (2021), "Utilizing a privacy-preserving IoT edge and fog architecture in automated household aquaponics", available at: http://www.ieomsociety.org/harare2020/papers/520.pdf

OpenMined/PyGrid (2021), "A peer-to-peer platform for secure, privacy-preserving, decentralized data science", GitHub, available at: https://github.com/OpenMined/PyGrid

Titre du rapport "Titre du rapport", available at: https://www.itu.int/en/ITU-D/Projects/ITU-EC-ACP/HIPSSA/Documents/FINAL%20DOCUMENTS/FINAL%20DOCS%20ENGLISH/sadc_model_law_data_protection.pdf

van der Sloot, B., Hoofnagle, C.J. and Zuiderveen, B.F.J. (2019), "The European Union general data protection regulation: what it is and what it means", *Information and Communications Technology Law*, Vol. 28 No. 1, pp. 65-98, doi: 10.1080/13600834.2019.1573501501.

Wolford, Ben (2020), "What is GDPR, the EU's new data protection law? 2020 - GDPR.eu. GDPR compliance", available at: https://gdpr.eu/what-is-gdpr/

Further reading

Alexander, A., Kajati, E., Koziorek, J. and Zolotova, I. (2022), "Federated learning for edge computing: a survey", available at: https://www.mdpi.com/2076-3417/12/18/9124/pdf-vor

Deconinck, K., Avery, E. and Jackson, L.A. (2021), Food supply chains and COVID-19 - impacts and policy lessons.

EU, GDPR "Privacy & security | Identification for development. ID4D", available at: https://id4d.worldbank.org/guide/privacy-security

FAO (2021), The state of food security and nutrition in the world.

Fronte, B., Galliano, G. and Bibbiani, C. (2016), From freshwater to marine aquaponic: new opportunities for marine fish species production.

Fung, C., Kadiyala, K., Jalali, F. and Dallas, U.T. (2019), "All one needs to know about fog computing and related edge computing paradigms: a complete survey all one needs to know about fog computing and related edge", *Computing Paradigms*.

Hevner, A.R. (2007), "A three cycle view of design science research", *Scandinavian Journal of Information Systems*, Vol. 19 No. 2, Article 4, available at: https://aisel.aisnet.org/sjis/vol19/iss2/4

Jiang, C., Qiu, Y., Gao, H., Li, K., Wan, J. and Fan, T. (2019), "An edge computing platform for intelligent operational monitoring in internet data centers", *IEEE Access*, Vol. 7, pp. 133375-133387, doi: 10.1109/ACCESS.2019.2939614.

Matrix Foundation (2014), "Matrix whitepaper", available at: https://www.matrixprotocol.io/whitepaper

Mohlameane, M. and Ruxwana, N. (2014), "The awareness of cloud computing: a case study of South African SMEs", *International Journal of Trade, Economics and Finance*, Vol. 5 No. 1, pp. 6-11, doi: 10.7763/IJTEF.2014.V5.332.

OpenMined/PyGrid (n.d), "*Practical law:* UK home", available at: https://uk.practicallaw.thomsonreuters.com/

Parikh, S., Dave, D., Patel, R., Doshi, N., Parikh, S., Dave, D., Patel, R. and Doshi, N. (2019), "ScienceDirect security and privacy issues in cloud, fog and edge computing security and privacy issues in cloud, fog and edge computing", *Procedia Computer Science*, Vol. 160, pp. 734-739, doi: 10.1016/j.procs.2019.11.018.

Peterson, Z.N.J., Gondree, M. and Beverly, R. (2014), A position paper on data sovereignty: the importance of geolocating data in the A position paper on data sovereignty: the importance of geolocating data in the cloud.

Profile, S.S.E.E. (2015), Aquaponics and its potential aquaculture wastewater treatment and human urine treatment Henrique Junior Aiveca Sánchez Licenciado em Ciências de Engenharia do Ambiente Aquaponics and its potential aquaculture wastewater treatment and human urine treatment Dissertação para obtenção do Grau de Mestre em.

Salman, O., Elhajj, I., Kayssi, A. and Chehab, A. (2020), Edge computing enabling the internet of Things.

Simangele, N. and Stephen, N. (2021), "AD371: limited Internet access in Zimbabwe a major hurdle for remote learning during pandemic", available at: https://afrobarometer.org/publications/ad371-limited-internet-access-zimbabwe-major-hurdle-remote-learning-during-pandemic

Wiewiórowski, W. (n.d), "D | European data protection supervisor. European data protection supervisor", available at: https://edps.europa.eu/data-protection/data-protection/glossary/d_en

World Vision (2020), "5 world hunger facts you need to know", available at: https://www.worldvision.org/hunger-news-stories/world-hunger-facts

Corresponding author
Solomon Hopewell Kembo can be contacted at: solomonkembo@gmail.com

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**