An Investigation into the Feasibility of Performing Federated Learning on Social Linked Data Servers

Nayil Arana
nra1u18@soton.ac.uk
University of Southampton
Southampton, UK

Mohamed Ragab
ragab.mohamed@soton.ac.uk
University of Southampton
Southampton, UK

Thanassis Tiropanis
t.tiropanis@soton.ac.uk
University of Southampton
Southampton, UK

ABSTRACT

Federated Learning (FL) and the Social Linked Data (`Solid` [1]) framework represent decentralized approaches to machine learning and web development, respectively, with a focus on preserving privacy. Federated learning enables the distributed training of machine learning models across datasets partitioned across multiple clients, whereas applications developed with the Solid approach store data in *Personal Online Data Stores* (pods) under the control of individual users. This paper discusses the merits and challenges of executing Federated Learning on Solid pods and the readiness of the Solid server architecture to support this. We aim to detail these challenges, in addition to identifying avenues for further work to fully harness the benefits of Federated Learning in Solid environments, where users retain sovereignty over their data.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Machine learning**; • **Security and privacy** → **Privacy-preserving protocols**.

KEYWORDS

Solid, Linked Data, Social Linked Data, pods, Federated Learning, Machine Learning, Privacy

ACM Reference Format:
Nayil Arana, Mohamed Ragab, and Thanassis Tiropanis. 2024. An Investigation into the Feasibility of Performing Federated Learning on Social Linked Data Servers. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3589335.3651950

1 INTRODUCTION

As tech giants increasingly monopolize personal data globally, data privacy and sovereignty concerns have intensified. Traditional data storage and machine learning methods rely on centralized data pools, posing significant ethical and legal risks, including privacy breaches under regulations such as the *GDPR* [4].

Researchers have proposed solutions like Federated Learning (FL) and Social Linked Data (Solid) to address these concerns. Federated Learning enables distributed model training across various clients without centralizing raw data, protecting privacy [1]. Solid is a web paradigm that aims to decentralise the web through the use of many user-controlled Personal Online Data Stores (pods) [2], as opposed to centralised data silos which do not promote data sovereignty [11].

As Solid applications gain adoption, analyzing data from Solid pods while preserving data privacy and sovereignty becomes crucial. Performing federated learning using client applications that access data from user Solid pods is one way to achieve this aim. In anticipation of this, this paper explores the viability of federated learning on Solid pods, laying the groundwork for future research.

2 BACKGROUND AND RELATED WORK

This section explores related work on Federated Learning and Solid, laying the foundation for a combined architecture.

McMahan et al. (2017) introduced Federated Learning through a baseline approach, Federated Averaging (FedAvg), establishing the foundation for distributed model training without central raw data aggregation [9]. FedAvg exemplifies a centralized (where a coordinating server merges local updates into a global model), horizontal (where data at each client is partitioned by samples) federated learning approach [6]. In this approach, clients update local models by stochastic gradient descent, with a central server averaging these updates [9].

The Solid platform, as conceptualised by *Mansour et al.*[8], is a web development paradigm that centers around Personal Online Data Stores (pods), which are online servers where each user in a system stores their data. Pods can be self-deployed on personal devices/servers or on cloud-based pod provider services [11]. Solid enables seamless switching between pod provider services, providing a competitive incentive for providers. Solid apps directly perform reads and writes on pods, enabling data reuse across different apps and decoupling application design from the data [11].

Solid also specifies protocols for accessing, reading from, and writing to resources stored on pods. These include WebID, which replaces traditional usernames and passwords with browser-stored profile information, the Resource Description Framework (RDF) for structured data storage, and the Linked Data Platform (LDP), which performs data manipulation through HTTP requests on resources defined by URIs [2].

[1] https://solidproject.org/

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651950

2.1 Previous Integration of FL and Solid - Research Gap

A review of the literature reveals a single attempt at integrating Solid and FL by Yuan et al [13], who designed a service recommendation system based on the use of a hybrid of horizontal and vertical federated learning on an extension of Solid.

While that work made the initial stride in combining FL and Solid, their focus was on a niche application, leaving the broader potential of using horizontal federated learning on Solid pods for more general classification or regression tasks unexplored. Furthermore, their work lacks details on system development and empirical speed evaluations [13].

3 APPLICATION ARCHITECTURE

This section proposes an architecture for performing FL on Solid pods across multiple users, illustrated in Fig. 1.

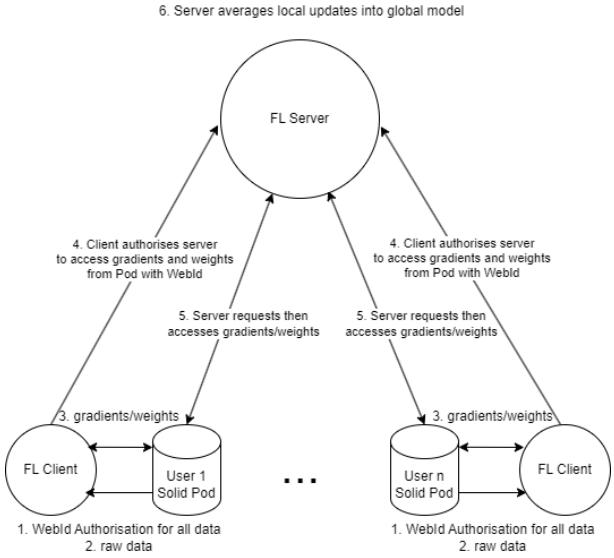

Figure 1: A diagrammatic representation of the proposed architecture for performing FL on Solid pods

As shown in Fig. 1, the process involves several steps enumerated below.

(1) The user authorises the FL client application to access all data from their pod by adding the client's WebID to the Pod's access control list (ACL)
(2) The FL client application accesses the user's raw data from the pod
(3) The client application performs local updates and stores the resulting gradients/weights in the user's pod
(4) The client adds the server's WebID to the pod's ACL, only authorising access to gradients and weights. The client also notifies the server to expect gradients/weights from the pod.
(5) The server requests, then accesses the gradients/weights from the user's pod.
(6) The server then averages the local updates into a global model.

By storing local model updates in users' pods and enabling the coordinating server to access only these updates, this architecture allows for multiple federated learning processes to be performed. Each process would have its own coordinating server, which accesses gradients/weights from pods that each user allows the server to have access to.

This type of architecture is worth exploring, as each pod owner may wish to participate in building one or more federated learning models based on different parts of the data in their pod. This is one way of ensuring sovereignty of the pod owners, as they can choose which FL projects they wish to participate in, without the need to provide any party with full access to their data.

4 DISCUSSION

In this section, we aim to explore the benefits and challenges of performing FL on Solid pods as envisioned in Section 3

FL naturally aligns with Solid's principles of data privacy and user sovereignty, as only model updates like gradients and weights are sent to the central server, allowing raw data to remain under user control without central aggregation. Furthermore, the decentralised nature of a system involving multiple pods potentially decreases the risk of a large-scale security breach - because of the difference in feasibility between separately attacking multiple pods, which may lie on different pod providers and user machines, and attacking a single centralised repository.

One challenge that must be addressed when performing FL on Solid pods is data format standardisation. This would be easy to manage in proof-of-concept projects, as splitting a publicly-available, clean dataset into pods was trivial. However, in a real-world system, users may not upload their data to the pod in a standardized format. Standardised protocols must be specified to ensure interoperability between data stored on different pods and FL systems.

A final challenge to mention is scalability. It can be deduced that accessing data from Solid pods introduces additional computational complexity to a federated learning system. This raises scalability concerns, as the additional execution time could become more pronounced as FL and Solid integration experiments grow in scale, with the eventual goal of deploying in production environments with many real users.

5 CONCLUSIONS AND FUTURE WORK

In summary, this paper achieved its objective of laying a foundation for future research in the area of performing FL on Solid pods. However, there is significant potential for further research in this area. Avenues for exploration are outlined in the below subsection.

5.1 Future Work

5.1.1 Domain-Specific Applications. Exploring the implementation of federated learning on Solid pods in specific domains in which this research is applicable could be particularly insightful, specifically in industries involving sensitive user data, namely healthcare [1], banking [7], and IOT [10]. Exploring the implementation of federated learning on Solid pods in specific domains in which this research is applicable could be particularly insightful. By applying FL-Solid integration research to these domains, one could discern

the domain-specific effectiveness of performing federated learning on Solid pods, as well as uncover challenges that may not be apparent in a more generalized approach.

5.1.2 Experimentation and Transition to Practical Application. Proof-of-concepts of the architecture proposed in Section 3 should be implemented in order evaluate its feasibility. After this, the culmination of this research would be in its real-world applicability. Transitioning from controlled simulations to a real-world scenario, where actual users, each equipped with their own Solid pod and a federated learning client app, could offer invaluable insights. This hands-on approach would further validate the scalability of federated learning across numerous Solid pods, and test the user-friendliness and feasibility of individual users managing their Solid pods to perform machine learning tasks. This transition would serve as a final demonstration of the practical effectiveness of the proposed system for performing FL on Solid pods.

5.1.3 Security Vulnerabilities. As FL and Solid advance towards real-world application, balancing technological progress with ethical considerations is crucial. Both aim to prioritize user privacy and data ownership, however they face potential security risks.

Federated learning is vulnerable to data poisoning and model inversion. Data poisoning refers to the injection of fabricated, malicious data into one of the clients, which will significantly decrease model accuracy [12]. Model inversion is an even greater concern, given how the proposed architecture detailed in Section 3 relies on sharing gradients and weights with the central server. It is possible to at least partially reconstruct data using gradients and weights, implying that sharing them may not be completely secure [5]. However, defenses against model inversion have been evaluated by the research community, including encrypting and adding noise to the gradients [5]. These defenses should be considered in future implementations of this research.

Solid faces its own set of security challenges. Solid does not intrinsically support access logs, which are needed for security and privacy auditing. Externalising this service to Pod providers and Solid applications exposes more privacy-sensitive interactions than necessary [3]. Additionally, the Solid paradigm does not specify any cryptographic means to protect data in a Pod other than HTTPS. This leaves potentially sensitive data vulnerable to attacks [3].

REFERENCES

[1] Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. 2022. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13 (2022), 1 – 23.

[2] Ruben Dedecker, Wout Slabbinck, Patrick Hochstenbach, Pieter Colpaert, and Ruben Verborgh. 2022. What's in a Pod?–A knowledge graph interpretation for the Solid ecosystem.

[3] Christian Esposito, Ross Horne, Livio Robaldo, Bart Buelens, and Elfi Goesaert. 2023. Assessing the Solid Protocol in Relation to Security and Privacy Obligations. *Information* 14, 7 (2023), 411.

[4] Kimberly A Houser and W Gregory Voss. 2018. GDPR: The end of Google and Facebook or a new paradigm in data privacy. *Rich. JL & Tech.* 25 (2018), 1.

[5] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. 2021. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 7232–7241.

[6] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, et al. 2021. Advances and Open Problems in Federated Learning. arXiv:cs.LG/1912.04977

[7] Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. 2021. Federated Learning for Open Banking. In *Federated Learning*.

[8] Essam Mansour, Andrei Vlad Sambra, Sandro Hawke, Maged Zereba, Sarven Capadisli, et al. 2016. A Demonstration of the Solid Platform for Social Web Applications. *Proceedings of the 25th International Conference Companion on World Wide Web* (2016).

[9] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[10] Dinh C. Nguyen, Ming Ding, Pubudu N. Pathirana, Aruna Seneviratne, Jun Li, et al. 2021. Federated Learning for Internet of Things: A Comprehensive Survey. *IEEE Communications Surveys and Tutorials* 23, 3 (1 July 2021), 1622–1658. https://doi.org/10.1109/COMST.2021.3075439

[11] Andrei Vlad Sambra, Essam Mansour, Sandro Hawke, Maged Zereba, Nicola Greco, et al. 2016. Solid: a platform for decentralized social applications based on linked data. *MIT CSAIL & Qatar Computing Research Institute, Tech. Rep.* (2016).

[12] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data poisoning attacks against federated learning systems. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*. Springer, 480–501.

[13] Haochen Yuan, Chao Ma, Zhenxiang Zhao, Xiaofei Xu, and Zhongjie Wang. 2022. A Privacy-Preserving Oriented Service Recommendation Approach based on Personal Data Cloud and Federated Learning. In *2022 IEEE International Conference on Web Services (ICWS)*. 322–330. https://doi.org/10.1109/ICWS55610.2022.00054