Towards a Blockchain-based Architecture for Data Provenance Management in the Internet of Things

Marcos Alves Vieira[1,2], **Sergio T. Carvalho**[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia Goiano (IF Goiano)
Iporá – GO – Brasil

[2]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brasil

marcos.vieira@ifgoiano.edu.br, sergiocarvalho@ufg.br

Abstract. *An Internet of Things (IoT) scenario is a heterogeneous and complex environment, where large volumes of data are constantly generated, manipulated, and transferred between different devices. In this context, some difficulties may arise, such as the correct identification of the devices generating the data, the trustworthiness of these devices and their generated data, detecting abnormal behavior, and controlling access to the data. Data provenance allows maintaining information about the origin of the data, the operations through which this data has undergone, and its processing history, from its creation to its current state. Aiming to provide means to mitigate the mentioned problems, we propose an architecture for data provenance management in IoT environments, enabling different levels of granularity, using a distributed ledger architecture.*

1. Introduction

The Internet of Things (IoT) enables interconnectivity between the digital and physical worlds, where sensors and actuators connect "things" (*e.g.*, objects, people, animals, machines, environments, infrastructure, vehicles) to each other and the Internet, making it possible to capture the context of these connected objects (*e.g.*, position) to provide services (*e.g.*, location-based services) [Groopman and Owyang 2018].

An IoT environment is naturally chaotic and heterogeneous, consisting of numerous and distinct devices that constantly exchange information with each other and with the Internet. In this context, several problems may arise, such as: *(a)* the correct identification of the devices generating the data (*e.g.*, "Did that information really come from that device?"; "Is the device that generated that information really whom it says it is?"); *(b)* the reliability of the devices and their generated data (*e.g.*, "Does this information make sense?"); *(c)* detection of abnormal behavior (*e.g.*, "This device has been generating discrepant data for some time; it is probably defective"); *(d)* controlling access to data (*e.g.*, ensuring that information is accessed only upon prior authorization).

Data provenance can provide means to solve the problems raised, by making it possible to keep a record of all the changes a piece of data goes through from its origin to its current state [Herschel et al. 2017]. Thus, when a problem occurs, it is possible to search the provenance history and trace the problem back to its origins. In [Hu et al. 2020], the authors state that in special cases, provenance data is more important than the original data itself. Therefore, when maintaining data provenance records, it must be stored

securely and reliably. Considering the distributed architecture of a blockchain network and its other characteristics, such as transparency, immutability, fault tolerance, and the absence of a central authority, this technology becomes very suitable to be used to store provenance data [Liang et al. 2017, Hu et al. 2020].

This paper presents the work in progress towards building an architecture that aims to provide means to mitigate the aforementioned problems, by providing data provenance management in IoT environments, making it possible to keep provenance records at different levels of granularity (*i.e.*, which provenance data will be recorded and how often), using as a basis a distributed ledger architecture in the form of a blockchain network, smart contracts and the W3C PROV language family[1].

The remainder of this paper is organized as follows: In Section 2, there is an analysis of the related works and a brief discussion about the differential of our proposal; Section 3 presents the theoretical and technological bases that this work relies on for its construction; Section 4 brings a breakdown of the proposed model, as well as an example of its application; Finally, in Section 5 the final considerations are outlined, including possibilities for future work.

2. Related Work

In this section, some works related to our proposal are presented, along with an analysis of their similarities and differences with the proposed architecture.

The work of [Margheri et al. 2020] presents a platform for managing the provenance of Electronic Health Records (EHRs), which can be implemented in already functioning EHR systems. The authors use blockchain technology as a basis, in addition to the Fast Healthcare Interoperability Resources (FHIR) to represent EHRs. A proxy transparently intercepts the EHRs' modifications and then triggers a smart contract to perform provenance annotations using the W3C PROV language. The resulting PROV document is stored in a Hyperledger Fabric blockchain.

In [Stoldt and Weber 2021], the authors dismiss the use of blockchain and perform provenance records directly on the patients' EHRs. They propose a reliability model aiming to assess the quality of medical data and support clinical decision-making. The method uses fuzzy logic to infer the level of reliability of the data produced, taking into account the level of trust of the data producers, the production method, and its certification. To this end, the authors performed an extension of the FHIR model to enable data provenance annotations to be stored directly in EHRs, making possible to verify the level of trustworthiness for each of the blood pressure records performed on the patients.

ProvChain [Liang et al. 2017] is an architecture to collect and verify data provenance in cloud computing. Similar to our proposal, data provenance is also stored in transactions in the blockchain. The operation of ProvChain is based on three phases: collection, storage, and validation of provenance data. A key component for ProvChain's operation is the *Provenance Auditor*, which has the role of retrieving and validating the information stored in the provenance database. The *Cloud User* represents the user who owns data and shares it with other users. *Provenance Database* is the database that reflects the state of the entire blockchain and is maintained by the Provenance Auditor.

[1] `https://www.w3.org/TR/prov-overview/`

Similar to our proposal, the work [Lautert et al. 2020] also provides a REST API to maintain provenance records on a blockchain, using the W3C PROV standard to manage the data provenance of a food supply chain. There are two main entities: Producers and Consumers of provenance data. The persistence of provenance data in the blockchain and its retrieval are in charge of the entity named *Provenance Service*.

The works mentioned in this section offer ways to manage data provenance for domain-specific systems (*e.g.*, cloud computing, electronic health records, and supply chain) using or not using blockchain technology. Our proposed architecture differs from these works as it can be used to maintain the data provenance of systems from different domains and at different levels of granularity by providing the developer with an Application Programming Interface (API) that can be easily used in their system. Thus, it is the developer's responsibility to decide which provenance data will be recorded, at which specific points in their application, and for what purpose (*e.g.*, auditing, access control, trustworthiness). It is worth noting that many authors (*e.g.*, [Dutta et al. 2020, Hu et al. 2020, Kumar et al. 2020]) argue that blockchain is a highly suitable solution for building data provenance systems. Additionally, by proposing an architecture to implement domain-independent provenance data management, our proposal addresses an open problem in the area of data provenance mentioned by [Hu et al. 2020].

3. Theoretical Background

This section brings the theoretical concepts on which this work is rooted. The operation of a blockchain was first described in [Nakamoto 2008] and consists of a structure of blocks that store data, where each block (except the first, named genesis block) stores the hash value of the previous block, forming a chain structure and ensuring the integrity of the entire blockchain.

The notion of a smart contract was introduced by [Szabo 1997] and allows a blockchain, other than storing states (data), to also store behaviors. Smart contracts are similar to contracts in the physical world, whose clauses describe actions that must be performed upon the occurrence of certain events. Through smart contracts, trust can be established between the parties without the need for a third party (*e.g.*, a notary), since the blockchain itself provides the guarantee that the smart contract will be executed as soon as its prerequisites are met.

The W3C Provenance Working Group[2] defines provenance as "information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness." The term originates from the French word "provenir", which means "coming from". Data provenance provides the history of the origins of all changes to an object, the list of components that have forwarded or processed that object, and the users that have viewed or changed it [Liang et al. 2017].

The W3C PROV is a standard which provides a means to represent provenance in the form of an XML schema, allowing, for instance, to store provenance data in a blockchain and take advantage of its inherent characteristics, such as: immutability, auditability, security, identity assurance, fault tolerance, lack of a central authority, among others [Shetty et al. 2017, Hu et al. 2020, Margheri et al. 2020].

[2] `https://www.w3.org/2011/prov/wiki/Main_Page`

4. Proposal

This section describes the proposed architecture aiming to maintain provenance records, based on a blockchain infrastructure, and also provides two examples of its application.

The concepts of an IoT architecture with data provenance support [Hu et al. 2020] are presented in Figure 1. The architecture we propose is located in the *Middleware Layer* and assumes that the other layers are defined and fully functioning.

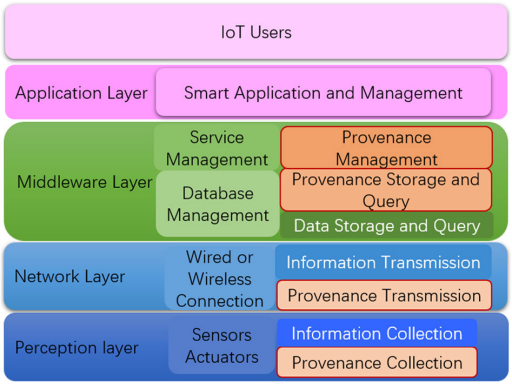

Figure 1. IoT architecture with support to data provenance [Hu et al. 2020].

4.1. IoT Architecture for Data Provenance in the Internet of Things

Figure 2 shows the components of the proposed architecture aiming to enable provenance management in IoT environments. Following, its components are detailed. It is worth noting that the proposal focuses only on the *Provenance Service Layer* and considers that the *IoT Application Layer* is under the responsibility of the application developer.

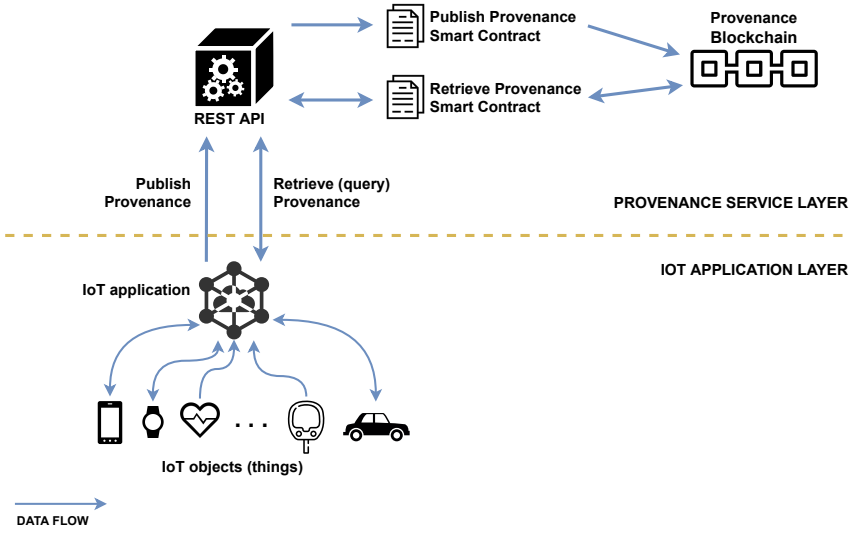

Figure 2. Architecture for data provenance management in IoT environments.

- **Rest API**: API designed to mediate the requests between the IoT Application and the smart contracts. It provides the IoT Application with operations to publish or to query the provenance data of a given object (thing) in the Provenance Blockchain.

- **Publish Provenance Smart Contract**: smart contract triggered by the Rest API, responsible for publishing provenance data to the Provenance Blockchain.
- **Retrieve Provenance Smart Contract**: smart contract triggered by Rest API, responsible for browsing the Provenance Blockchain for provenance data, based on the given search criteria, and returning it to the Rest API, which in turn forwards it to the IoT Application.
- **Provenance Blockchain**: blockchain infrastructure where provenance data is stored in the form of blocks, following the W3C PROV standard.

The proposed architecture aims to facilitate the management of provenance data for the IoT application developer. Therefore, the developer can simply perform calls to the Rest API in order to publish or retrieve provenance data about a given object (thing) that is stored in the Provenance Blockchain. It remains with the developer to decide the level of granularity of the provenance that will be maintained (*i.e.*, which provenance data will be recorded and how often), as well as its purpose, which can be, for instance, to enable auditing over the application in case of failures or abnormal behavior, to ensure the reliability of the objects and their produced data, or to maintain access control over the information generated by the application.

4.2. Applicability Example

In [Vieira and Carvalho 2016a, Vieira and Carvalho 2016b], we present a scenario of a smart home inhabited by an elderly couple who use wearable medical devices to monitor their vital signs. Other IoT devices are deployed throughout the house, such as a digital scale, motion sensors, and smart locks. All of these IoT devices are connected to a personal health application that aims to monitor the couple's vital signs and send them reminders to take their medications on time, perform regular physical activity, or check their weight on the digital scale.

The provenance management architecture proposed in this paper would be useful to keep provenance records of the IoT devices in this smart home in order to, for instance: *(a)* ensure the identity of the devices generating the data; *(b)* identify the exact moment when a faulty device started reporting inconsistent data and ignore its measurements thereafter; *(c)* enable patients to perform third-party concession to their monitored data.

5. Concluding Remarks

The heterogeneity of IoT environments and the constant generation and exchange of data between different devices that constitute it can lead to several challenges, such as device reliability and correct identification. Data provenance makes it possible to track all the changes a piece of data undergoes, from the time it is created to its current state. Thus, maintaining the data provenance of a system enables diagnosing problems, assessing the quality of data, ensuring the trustworthiness of a device, and so on.

This paper presented ongoing work towards building an architecture to be used by developers to store domain-independent provenance data at different levels of granularity, supported on top of a blockchain infrastructure and using the W3C PROV provenance representation standard.

The next steps include testing in simulated environments from different domains, such as those mentioned in Section 4.2, as a way to validate the architecture and ensure

that it is able to handle provenance data at different levels of granularity. An important outcome of this work is to assess the pros and cons of combining blockchain and cloud computing technologies to store the provenance data in the cloud (off-chain) and keep in the blockchain only a reference for the data that is in the cloud, along with its hash value, as a way to ensure the integrity of the provenance data.

References

[Dutta et al. 2020] Dutta, P., Choi, T.-M., Somani, S., and Butala, R. (2020). Blockchain technology in supply chain operations: Applications, challenges and research opportunities. *Transportation Research Part E: Logistics and Transportation Review*, 142:102067.

[Groopman and Owyang 2018] Groopman, J. and Owyang, J. (2018). The Internet of Trusted Things. *Kaleido Insights: The Internet of Trusted Things: Blockchain as the Foundation for Autonomous Products & Ecosystem Services*, page 22.

[Herschel et al. 2017] Herschel, M., Diestelkämper, R., and Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? *The VLDB Journal*, 26(6):881–906.

[Hu et al. 2020] Hu, R., Yan, Z., Ding, W., and Yang, L. T. (2020). A survey on data provenance in IoT. *World Wide Web*, 23(2):1441–1463.

[Kumar et al. 2020] Kumar, A., Liu, R., and Shan, Z. (2020). Is blockchain a silver bullet for supply chain management? *Decision Sciences*, 51(1):8–37.

[Lautert et al. 2020] Lautert, F., Pigatto, D. F. G., and Gomes-JR, L. C. (2020). Blockchain-based Data Provenance. In *III Workshop em Blockchain: Teoria, Tecnologias e Aplicações (WBlockchain 2020)*, Rio de Janeiro – RJ – Brazil. Sociedade Brasileira de Computação.

[Liang et al. 2017] Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., and Njilla, L. (2017). ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, pages 468–477. IEEE.

[Margheri et al. 2020] Margheri, A., Masi, M., Miladi, A., Sassone, V., and Rosenzweig, J. (2020). Decentralised provenance for healthcare data. *International Journal of Medical Informatics*, 141(June):104197.

[Nakamoto 2008] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. White Paper. Available online: `https://bitcoin.org/bitcoin.pdf`.

[Shetty et al. 2017] Shetty, S., Red, V., Kamhoua, C., Kwiat, K., and Njilla, L. (2017). Data provenance assurance in the cloud using blockchain. In Hall, R. D., Blowers, M., and Williams, J., editors, *Disruptive Technologies in Sensors and Sensor Systems*, volume 10206, page 102060I.

[Stoldt and Weber 2021] Stoldt, J.-P. and Weber, J. H. (2021). Provenance-based Trust Model for Assessing Data Quality during Clinical Decision Making. In *3rd ICSE Workshop on Software Engineering for Healthcare*.

[Szabo 1997] Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday*, 2(9).

[Vieira and Carvalho 2016a] Vieira, M. A. and Carvalho, S. T. (2016a). Addressing the concurrent access to smart objects in ubiquitous computing scenarios. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, page 79–82, New York, NY, USA. Association for Computing Machinery.

[Vieira and Carvalho 2016b] Vieira, M. A. and Carvalho, S. T. (2016b). (Meta)Modelagem de Espaços Inteligentes Pessoais e Espaços Inteligentes Fixos para Aplicações Ubíquas. In *Anais do VIII Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)*, pages 1056–1065, Porto Alegre-RS, Brazil. Sociedade Brasileira de Computação - SBC.