

Blockchain-based Data Provenance for the Internet of Things

Marten Sigwart

TU Wien, Vienna, Austria
m.sigwart@infosys.tuwien.ac.at

Michael Borkowski

TU Wien, Vienna, Austria
m.borkowski@infosys.tuwien.ac.at

Marco Peise

TU Berlin, Berlin, Germany
mp@ise.tu-berlin.de

Stefan Schulte

TU Wien, Vienna, Austria
s.schulte@infosys.tuwien.ac.at

Stefan Tai

TU Berlin, Berlin, Germany
st@ise.tu-berlin.de

ABSTRACT

As more and more applications and services depend on data collected and provided by Internet of Things (IoT) devices, it is of importance that such data can be trusted. Data provenance solutions together with blockchain technology are one way to make data more trustworthy. However, current solutions do not address the heterogeneous nature of IoT applications and their data.

In this work, we identify functional and non-functional requirements for a generic IoT data provenance framework, and conceptualise the framework as a layered architecture. Using a proof-of-concept implementation based on Ethereum smart contracts, data provenance can be realised for a wide range of IoT use cases. Benefits of a generic framework include simplified adoption and a more rapid implementation of data provenance for the IoT.

ACM Reference Format:

Marten Sigwart, Michael Borkowski, Marco Peise, Stefan Schulte, and Stefan Tai. 2019. Blockchain-based Data Provenance for the Internet of Things. In *Proceedings of 9th International Conference on the Internet of Things (IOT'19)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The Internet of Things (IoT) is transforming many areas of our everyday lives. IoT technologies such as GPS, RFID-based identification, and low-resource computing platforms like

the Raspberry Pi already play important roles in various domains, e.g., mobility, logistics, healthcare, and retail [1]. Since data collected by these technologies find application in an increasing number of use cases, ensuring the trustworthiness of such data is of high importance [2].

Data provenance systems are one way to ensure trustworthiness in the IoT [3]. These systems can provide information about the origin and evolution of data such as the various stages of data creation and data modifications, who initiated them, and when and how they took place [4]. In short, a data provenance system tracks who has created, updated, deleted, and in some cases read particular data points [4]. In order to trust the information provided by a provenance system, it is essential that the provenance system stores information in a tamper-proof and replicable way [5].

Traditionally, in distributed settings, participants must either trust each other or an independent third-party to store data in a tamper-proof way. With the advent of blockchain technologies, the requirement of such trust in a central authority is eliminated. Instead, a decentralised network can be established, acting as a distributed, tamper-proof ledger [6]. Thus, leveraging blockchain technology in a data provenance solution for the IoT is a promising choice [6].

The IoT is characterised by a multitude of use cases and potential application areas [1]. An IoT data provenance solution has to account for this diversity [7]. However, approaches aiming at scenario-agnostic blockchain-based data provenance solutions for the IoT offer so far no concrete software solutions [8, 9], and more concrete solutions to data provenance in the IoT only focus on specific application areas, for instance, supply chains [10–12], health monitoring systems [13], or digital forensics [14].

Hence, we propose a generic blockchain-based data provenance framework for the IoT that can be applied to a variety of use cases. The advantages of a generic framework are the easier adoption of provenance concepts by new use cases and the interoperability of applications that use the framework. The contributions of this work are as follows:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IOT'19, October 22–25, Bilbao, Spain

© 2019 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

- We define functional and non-functional requirements for a generic IoT data provenance framework.
- We conceptualize and implement an IoT data provenance framework consisting of smart contracts using a generic data model to provide provenance functionality for a wide range of IoT use cases.
- We present an evaluation of the framework with regard to the defined requirements using a proof-of-concept implementation with Ethereum smart contracts.

Following, we briefly explain underlying concepts and provide exemplary use case scenarios (Section 2), define the requirements (Section 3), and explain the concepts, architecture, and proof-of-concept implementation of our IoT data provenance framework (Section 4). In Section 5, we evaluate the framework with regard to the defined requirements. Section 6 provides an overview of the related work. Finally, Section 7 concludes the paper.

2 BACKGROUND & MOTIVATION

Data provenance, sometimes also known as lineage or pedigree [15], identifies the derivation history of data [4]. While originally used for works of art, data provenance is now relevant in a wide range of use cases, since data provenance mechanisms help to establish a certain level of trust in data by providing information about its creation, access, and transfer [4]. Provided provenance data is secured, forgery, alteration or repudiation of data can be prevented [5]. Accordingly, data provenance solutions can help to establish trust in data in the IoT [2]. In the following, we describe two exemplary use cases for data provenance in the IoT.

Vaccine Supply Chains. Immunisation programmes depend on functional, end-to-end supply chains [16]. In all phases of the supply chain—from procurement to last-mile distribution—it is of significant importance that vaccines remain in a temperature range of around 2–8 °C. Otherwise, vaccines lose their effectiveness. An unbroken, temperature-controlled link from producer to consumer is also referred to as the cold chain [16]. In IoT-enabled cold chains, technologies such as RFID, GPS and sensing technologies [1] track temperatures, locations and other conditions of the vaccines. This provenance data helps to establish confidence in the quality of vaccines and exposes any weak links along the supply chain. As mentioned in Section 1, an important requirement for provenance systems is that recorded provenance information is replicable and tamper-proof. Recent scandals¹ of vaccine counterfeiting have confirmed the urgency of these requirements in vaccine supply chain systems. Hence, an IoT-powered data provenance system based on blockchain

technology could provide benefits in vaccine supply chain scenarios. In particular, it enables the backtracking of errors in case of breakage of the cold chain, and it helps to build and to keep trust in immunisation programmes by preventing counterfeit vaccines from entering the supply chain.

Health Monitoring Systems. Health monitoring scenarios are another application area where tamper-proof data provenance records can provide additional trust in data [13]. In such scenarios, sensors monitor health conditions of patients such as a patient's heart rate, blood pressure, etc. Combined with a real-time analytics service, the sensor readings can be used to notify relatives and health professionals in case of medical emergencies, such as a heart attack. Ideally, the root cause of such a notification is verifiable, i.e., it is possible to trace back a notification to the individual sensor readings which caused it. Otherwise, if the circumstances that led to the emergency notification are unclear, the possibility that the notification was caused by a malicious attack tampering with the system cannot be excluded. Having tamper-proof provenance data for a notification, such as information about the individual sensor readings and analytics involved, ensures the reliability and accuracy of the system.

3 REQUIREMENTS

This section defines the functional and non-functional requirements of a generic data provenance framework for the IoT. Such a framework needs to record provenance information for addressable data points, i.e., data that has a unique ID [7]. In accordance with [4] and [7], we derive the functional requirements (Req. 1–7) of our framework. Further, in accordance with Hasan et al. [5], we define non-functional requirements that need to be fulfilled by a tamper-proof data provenance system for the IoT (Req. 8–11).

- (1) *Provenance Abstraction:* The framework needs to provide generic data provenance capturing, storing, and querying functionality which can be adopted by provenance use cases to map their specific requirements.
- (2) *High-level and Low-level Provenance:* Provenance records represent high-level as well as low-level data points. Low-level data points stem from low-level devices such as sensors. High-level data points do not have a single physical origin (such as a sensor reading) but represent more abstract concepts, e.g., some physical object in a supply chain or an analytics result based on multiple inputs.
- (3) *Completeness:* A provenance record is complete if every relevant action which has ever been performed on a data point is gathered [5]. Here, relevance implies that some actions can be neglected if they do not contribute to the provenance information of a particular data point.
- (4) *Creation of Lineage:* Provenance records for a data point can be created based on the last provenance record for

¹<https://www.securindustry.com/pharmaceuticals/massive-fake-vaccine-racket-busted-in-indonesia/s40/a2849/#>.

that same data point. This enables the creation of lineage. For instance, the lineage of a data point representing a physical good travelling along a supply chain can be tracked by creating a new provenance record based on the old one at each critical step of the supply chain.

- (5) *Derivation*: A provenance record entails references to the provenance records of the data points that led to its creation. For instance, a provenance record for an analytics result based on value readings from multiple sensors must not only contain information about the sensor values but must also be able to access provenance information of those same values, such as location and time of recording.
- (6) *Provenance for Modifications of Data Points*: The framework enables the tracking of the modification history of a specific data point. For instance, if an analytics result runs through multiple stages of different calculations, the history of these calculations can be tracked.
- (7) *Parallel Provenance*: Multiple provenance records for the same data point can exist in parallel, e.g., one provenance trace might track the ownership of a data point (e.g., ownership of a physical good), while another provenance record is tracking the location of that same data point.
- (8) *Integrity*: Integrity mandates that provenance records cannot be manipulated or modified by an adversary in any way. This is crucial for establishing trust in the data. Without guaranteed integrity, clients can potentially repudiate provenance records.
- (9) *Availability*: Data can be accessed when needed.
- (10) *Privacy*: Provenance records of IoT devices can contain sensitive data, e.g., in a health monitoring system. It is therefore vital to keep this data confidential, i.e., to prevent access by unauthorised entities. In addition, even when the data itself is kept confidential, malicious actors might still be able to create user profiles by identifying user-specific data patterns. Hence, traceability of provenance data needs to be prevented to ensure user privacy.
- (11) *Scalability*: A provenance system is required to have a reasonable expenditure. Storing and accessing provenance information must have a low overhead. Especially, even though some IoT devices are resource-constrained, they must not be excluded from participating in the provenance framework. Further, applications in the IoT potentially deal with massive amounts of data and very frequent data updates which a provenance solution needs to account for.

4 IOT DATA PROVENANCE FRAMEWORK

This section presents the proposed generic blockchain-based data provenance framework for the IoT. As mentioned in Section 1, the benefits of a generic framework are interoperability between applications using the framework and the

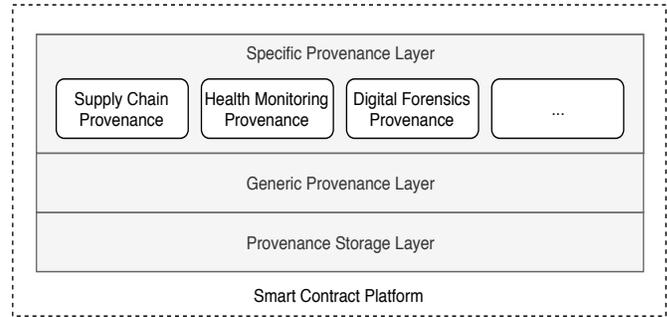


Figure 1: Data Provenance Framework Architecture

facilitated implementation of provenance concepts for new IoT use cases. Figure 1 displays the core architecture of the framework. It consists of three layers all embedded within a blockchain smart contract platform. Each layer represents a different level of abstraction with a diverse set of responsibilities within the framework. The storage layer is primarily concerned with low-level representation and storage of provenance data, the generic provenance layer provides general-purpose provenance functionality, while the specific provenance layer can be modified by use cases to fine-tune the framework to their specific requirements.

A proof-of-concept implementation of the framework based on Ethereum smart contracts is available as open-source software at Github².

Data Model

Provenance information varies depending on the specific domain or application [7]. Since the IoT domain is characterised by heterogeneous applications and a wide range of possible use cases [1], there is a need for a generalised provenance model suitable for IoT applications. As the underlying data model, our framework utilises the data provenance model by Olufowobi et al. [7] since their model is specifically designed to address IoT provenance data. The model defines a provenance record for some data point dp as follows:

$$prov : dp \mapsto \langle addr(dp), \{prov(idp) \mid \forall idp \in inputs(dp)\}, context(dp) \rangle$$

A provenance record $prov(dp)$ for a data point dp consists of a 3-tuple associating the address $addr(dp)$ of dp (i.e., some kind of ID) with the set of provenance records of the data points that led to the creation of dp (i.e., $inputs(dp)$), and a context $context(dp)$. The context denotes information of interest for provenance about the state of the IoT system, e.g., information about agents involved in the computation of the data point, time and location, or the execution context. This definition of provenance allows for the description of both

²<https://github.com/msigwart/iotprovenance>

Listing 1: Storage Contract (Excerpt)

```

1 contract Storage {
2   struct ProvenanceRecord {
3     uint tokenId;
4     uint[] inputProvenanceIds;
5     string context;
6     uint index;
7   }
8   mapping (uint -> ProvenanceRecord) records;
9   uint[] provenanceIndex;
10  function getProvenance(uint provId)
11  public view returns (uint tokenId, ...) {...}
12  ...
13  function createProvenance(uint provId, ...)
14  internal returns (uint index) {...}
15  ...
16 }

```

creation and modification of data points. In the former case, the set of input provenance records contains an empty set. In the latter case, the set of input provenance records contains the provenance record of the data point before modification, i.e., $prov(dp') = \langle addr(dp'), \{prov(dp), \dots\}, context(dp') \rangle$.

The described data provenance model combined with blockchain technology builds the basis for our IoT data provenance framework. This way, the framework can not only represent provenance information for various IoT use cases, but also provides integrity guarantees for the recorded data.

Storage Layer

This layer is responsible for the low-level storage of provenance records. It contains the generic representation for provenance records as defined by the data provenance model, as well as basic functionality to create, retrieve, update, and delete provenance records. Delete refers to the invalidation of provenance records, since truly deleting data from a blockchain is not possible. An invalidated provenance record cannot be used as input for subsequent provenance records.

Listing 1 displays an excerpt of the smart contract implementing this layer. The internal representation of provenance records (Lines 2–7) closely resembles the model discussed above with the field *tokenId* representing the ID of a data point. However, not only are data points addressable, but also the provenance records themselves. Addressable provenance records allow the storage layer to manage provenance records as a mapping from provenance IDs to provenance records: $addr(prov) \mapsto prov$ where the function $addr(prov)$ represents the ID of a provenance record *prov*. The mapping is implemented via the *mapping* keyword (Line 8).

The contract exposes an API for creating, retrieving, updating, and deleting (i.e., invalidating) provenance records. Note that, while functionality for retrieving provenance records

is exposed publicly via the *public* keyword (Lines 10ff), functionality for creating, updating, and deleting records is protected via the *internal* keyword (Lines 13ff). These functions cannot be accessed publicly, but are accessible from inheriting contracts such as the contract representing the generic provenance layer. Read-functions are publicly accessible since these functions do not alter the state of the contract.

Generic Provenance Layer

The generic provenance layer's main purpose is to provide general-purpose provenance functionality on top of the storage layer, i.e., it provides features that are universally applicable for a wide range of provenance use cases. The generic provenance layer has the following responsibilities.

Ownership of Data Points. While blockchain technology can guarantee the integrity of data provenance records once those records have entered the system, mechanisms need to be in place to ensure that the records that enter the system are correct. As a first step, we aim to prevent the creation of provenance records by arbitrary clients, i.e., if client *A* generates some data point dp_0 , we want to make sure that only client *A* (or any client authorised by client *A*) is able to create provenance records for dp_0 . Thus, we introduce the notion of ownership of data points. Each data point belongs to a specific client of the system, and only the owner or a client authorised by the owner can create provenance records for it. If an unauthorised client tries to create a provenance record for a data point, the system raises an error.

The notion of ownership is closely related to so-called tokens, i.e., smart contracts deployed on a blockchain that represent a kind of digital asset [17]. Our framework leverages tokens to introduce ownership of data points. Each data point is represented by a single token and a single token identifies exactly one data point. This one-to-one mapping allows us to identify the owner of a data point by identifying the owner of a particular token.

Each token acts as an entry ticket to the provenance framework. To create provenance records for a particular data point, a client first has to become the owner or be approved by the owner of the corresponding token. The prototype uses the Ethereum token standard ERC721³ which defines a common interface for non-fungible assets, for instance, functions for transferring ownership. This has the advantage that our tokens (i.e., the data points) can be traded by any client implementing this standard, such as wallets or exchanges. By transferring ownership, data points can pass from owner to owner, leaving a trail of provenance records created by each owner along the way. This is useful for implementing provenance applications, e.g., in supply chain scenarios.

³<https://github.com/ethereum/EIPs/blob/master/EIPS/eip-721.md>

Listing 2: Generic Provenance Contract (Excerpt)

```

1 contract GenericProvenance is ERC721, Storage {
2 ...
3 mapping (uint => uint[]) internal associatedProv;
4 ...
5 function createProvenance(uint tokenId, ...)
6 internal returns (uint index) {
7     require(exists(tokenId));
8     require(ownerOf(tokenId) == msg.sender);
9     checkValidProvenance(inputProvenance);
10    uint provId = getProvId();
11    addAssociatedProvenance(tokenId, provId);
12    return super.createProvenance(provId, ...);
13 }
14 ...

```

Associating Provenance Records with Data Points. The generic provenance layer further links together data points and their respective provenance records. The framework provides information about associated provenance records of specific data points. Within the generic provenance layer, this is achieved by using a mapping from a data point ID to a set of associated provenance IDs:

$$addr(dp) \mapsto \{addr(prov_1(dp)), addr(prov_2(dp)), \dots\}$$

All associated provenance records ($prov_1$, $prov_2$, etc.) represent parallel provenance traces for the same data point. Of those records, each one represents a completely independent trail of provenance information. For instance, one trail of provenance records might trace the temperature history of a physical good, while another one might trace its location.

Listing 2 displays an excerpt of the smart contract implementing this layer and demonstrates the general workflow for creating new provenance records. First, the contract verifies that a given token (i.e., data point) exists (Line 7) and that the token belongs to the sender of the message (Line 8). After validating the input provenance records (Line 9), the contract creates a new provenance ID (Line 10), adds it to the list of associated provenance (Line 11), and calls the storage contract’s *createProvenance* function (Line 12). Note that the *createProvenance* function of the generic contract is again internal, and thus not accessible publicly. This enables specific provenance contracts which implement concrete use cases to further adapt the functionality according to their needs.

Specific Provenance Layer

Smart contracts within this layer utilise the functionality provided by the generic provenance layer, but control a subset of parameters by themselves. This way, use cases can customise the provenance model according to their needs, and control access to the functionality provided by the storage

Listing 3: Specific Provenance Contract (Example)

```

1 contract Specific is GenericProvenance {
2 function requestToken()
3 public returns (uint tokenId) {
4     uint tokenId = getTokenId();
5     return super.mint(msg.sender, tokenId);
6 }
7 function createProvenance(uint tokenId,
8 string location, uint temperature, ...)
9 public returns (uint index) {
10    string context = createContext(location,
11    temperature, ...);
12    return super.createProvenance(provId,
13    context, ...);
14 }
15 ...

```

and generic provenance layer. Listing 3 displays an excerpt of an exemplary smart contract implementing this layer.

The provenance model can be customised by defining the *context* parameter, so that it presents the provenance information needed for a specific use case scenario. For instance, in the case of vaccine supply chains, the context should contain temperature and location information about individual vaccines. Hence, a specific contract could define its own *createProvenance* function that requires parameters like *temperature*, and *location* (Lines 7ff) which are then combined to form the *context* to be passed on to the *createProvenance* function of the generic provenance contract (Lines 10ff).

Access control happens on two levels. First, a specific contract defines which parts of the generic provenance layer’s API are exposed. For instance, even though the generic provenance layer could permit updating or deleting (i.e., invalidating) provenance records, this could be unwanted behaviour in the specific use case at hand. In this case, the contract in the specific provenance layer simply “hides” the functionality, i.e., does not expose it publicly.

Second, access control is relevant for controlling the ownership of data points. Since data point ownership is the decisive factor with regard to who can create provenance records for which particular data points, contracts in the specific provenance layer are responsible for actually assigning ownership, i.e., which tokens get assigned to which clients. For simplicity, our prototype automatically assigns new tokens to requesting clients (Lines 2ff). However, ultimately, the most suitable approach is dictated by the specific use case at hand. For instance, clients could purchase tokens. This would keep the framework publicly available while reducing the risk of spamming attacks since the acquisition of tokens incurs financial cost. Another alternative is a white list of authorised clients allowed to request new tokens. This way, the list controls exactly who participates in the provenance

system. However, the question arises who is responsible for managing the white list of authorised clients.

5 EVALUATION

We evaluate the framework with regard to the functional and non-functional requirements defined in Section 3. The use cases defined in Section 2 act as basis for the evaluation of Req. 1–3. In addition, we reason about the fulfilment of Req. 4–7 in an exemplary fashion applying the use case of vaccine supply chains. However, the scenario of vaccine supply chains is merely used to provide a more descriptive analysis. The information specific to the vaccine supply chain can be substituted with data reflecting any other use case. While the presented framework is blockchain-agnostic, the non-functional requirements (Req. 8–11) are evaluated using the proof-of-concept implementation. Experiments are performed on the public Ethereum test networks Rinkeby⁴ and Ropsten⁵. Rinkeby and Ropsten are chosen as test networks since their average block times most closely resemble the block times of the main Ethereum network.

Provenance Abstraction (Req. 1). The presented framework consists of multiple abstraction layers. The storage layer is responsible for low-level storage of provenance records while the generic provenance layer extends the storage layer's functionality with generic provenance features and enhanced access control. The generic provenance layer can be extended by use case-specific smart contracts controlling certain parameters of the application, e.g., by assigning ownership of tokens, and/or by exposing or hiding parts of the generic layer's API. For instance, in the scenario of vaccine supply chains, a specific provenance smart contract might give an entity such as the World Health Organisation complete control over who is able to acquire tokens, e.g., only trusted vaccine manufacturers. In the use case of health monitoring systems, multiple approved manufacturers might have control over assigning ownership of data point IDs. Hence, we conclude that the framework acts as a base abstraction which can be extended to fulfil the needs of specific provenance use cases. Thus, we regard Req. 1 as fulfilled.

High-level and Low-level Provenance (Req. 2). The framework identifies each data point by its corresponding token. As long as a unique token (i.e., ID) gets assigned to a data point, provenance information for that data point can be recorded. Hence, the framework does not pose any restrictions on the nature of the data point. It is possible to record provenance information for high-level as well as low-level data points. The *context* parameter of a provenance record can be used to add any kind of information the user desires.

In the use case of vaccine supply chains, provenance information could not only be collected for the vaccines themselves, but also for sensor readings along the supply chain, e.g., temperature readings which document an uninterrupted cold chain. Within health monitoring systems, the patients themselves might be represented by data points. The patients' provenance traces are then augmented by provenance information from low-level data points deriving from medical devices. Hence, we regard Req. 2 as fulfilled.

Completeness (Req. 3). The broad definition of the *context* parameter allows the collection of all information necessary for complete provenance tracking. The definition of completeness is largely dependent on the specific provenance use case. In the example of vaccine supply chains, information required to prove the origin of vaccines and an uninterrupted cold chain can be recorded. Regarding health monitoring systems, we are able to record all information necessary to provide reliable root cause information for medical emergencies. Therefore, Req. 3 can also be regarded as fulfilled.

Creation of Lineage (Req. 4). As an example, we assume that the manufacturer *SaferVaccines Inc.* produces a new vaccine $vacc_1$. The freshly produced vaccine is packaged with an RFID tag and assigned with a unique ID. A completely new provenance record for the vaccine is created, such as $prov(vacc_1) = \langle addr(vacc_1), \emptyset, \langle agent=operator_1@SaferVaccinesInc, time=5am, \dots \rangle \rangle$. Vaccine $vacc_1$ is loaded onto an aircraft air_1 . An RFID reader at the factory gate scans the vaccine and registers a new provenance record of the vaccine leaving the factory, such as $prov(vacc_1)' = \langle addr(vacc_1), \{prov(vacc_1)\}, \langle agent=rfid_1@SaferVaccinesInc, time=5am, \dots \rangle \rangle$. Shortly after, an RFID reader at the aircraft's entrance registers the vaccine entering the aircraft, e.g., $prov(vacc_1)'' = \langle addr(vacc_1), \{prov(vacc_1)'\}, \langle agent=rfid_1@air_1, time=5am, \dots \rangle \rangle$. The provenance records $prov(vacc_1)$, $prov(vacc_1)'$, and $prov(vacc_1)''$ represent the lineage of the vaccine. This shows exemplarily how the framework fulfils Req. 4.

Derivation (Req. 5). Continuing the example from Req. 4, inside the aircraft, sensors constantly monitor the temperature. There are three readings from a temperature sensor: $dp_1 = 38^\circ\text{F}$, $dp_2 = 45^\circ\text{F}$, $dp_3 = 40^\circ\text{F}$. The provenance records are given as follows:

$$\begin{aligned} prov(dp_1) &= \langle addr(dp_1), \emptyset, \langle agent=sensor, time=7am, \dots \rangle \rangle \\ prov(dp_2) &= \langle addr(dp_2), \emptyset, \langle agent=sensor, time=8am, \dots \rangle \rangle \\ prov(dp_3) &= \langle addr(dp_3), \emptyset, \langle agent=sensor, time=9am, \dots \rangle \rangle \end{aligned}$$

A fourth data point dp_4 is created by a software agent calculating the average temperature, i.e., $dp_4 = (dp_1 + dp_2 + dp_3)/3 = 41^\circ\text{F}$. This data point's provenance record is defined as $prov(dp_4) = \langle addr(dp_4), \{prov(dp_1), prov(dp_2), prov(dp_3)\}, \langle agent=averager@air_1, time=10am, \dots \rangle \rangle$. Hence, the framework allows for the creation of provenance records for data points deriving from multiple other data points (Req. 5).

⁴<https://rinkeby.etherscan.io/>

⁵<https://ropsten.etherscan.io/>

Provenance for Modifications of Data Points (Req. 6). In a further step, the calculated average temperature dp_4 is converted into a different unit: $dp'_4 = 5^\circ\text{C}$ (i.e., from Fahrenheit to Celsius). The resulting provenance record looks like $prov(dp'_4) = \langle addr(dp'_4), \{prov(dp_4)\}, \langle agent=converter@air_1, time=11am, \dots \rangle \rangle$. Thus, the framework is also able to support the modification of data points (Req. 6).

Parallel Provenance (Req. 7). Besides measuring the temperature inside the aircraft (air_1), we might also want to track the location of the aircraft at the same time. The framework enables the creation of parallel provenance records since each data point is mapped to a list of associated provenance records (see Section 4). We can create one provenance record $prov_{temperature}(air_1)$ representing the latest temperature inside the aircraft, and one provenance record $prov_{location}(air_1)$ representing the latest location of the aircraft. Hence, we regard Req. 7 as fulfilled.

Integrity (Req. 8). The presented framework uses blockchain technology to provide trustless and tamper-proof storage of provenance records for IoT data. Thus, once records have entered the system, the integrity of records depends on the underlying blockchain technology. Ethereum is a public blockchain with around 8250 fully validating nodes securing the network at the time of writing⁶. Hence, we consider the integrity requirement for data and computations on Ethereum as fulfilled. However, the framework also needs to provide mechanisms to ensure only correct records enter the system in the first place. As a first step, we implement the concept of ownership of data points using tokens to prevent arbitrary clients from creating provenance records. In future work, this concept could be further extended by mechanisms that guarantee the correctness of the records themselves.

Availability (Req. 9). In public blockchain networks like Ethereum potentially anyone can run a full node. Hence, we consider the availability requirement as fulfilled.

Privacy (Req. 10). While the privacy requirements of a provenance system largely depend on the concrete use case at hand, in scenarios with sensitive data, such as a health monitoring system, privacy is crucial [13]. However, privacy remains an ongoing challenge in the realm of (public) blockchains, since a blockchain’s security relies on data being transparent and verifiable by every participant. A blockchain-based data provenance framework suffers from the same problem. While some propose to use private blockchains in scenarios where privacy is important [13], this is not sufficient in scenarios which also depend on the system being publicly available, such as global vaccine supply chains.

Scalability (Req. 11). Since IoT scenarios potentially deal with massive amounts of data and frequent data updates,

IOT’19, October 22–25, Bilbao, Spain

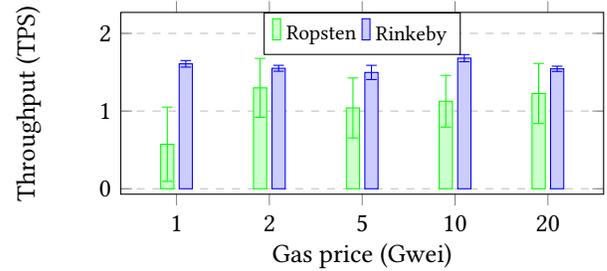


Figure 2: Avg. Throughput for Creating Provenance Records

we evaluate the framework’s scalability with regard to latency and transaction throughput. Latency and throughput are mostly influenced by two factors, network load (how many transactions are submitted to the blockchain) and gas price (how much are the senders of the individual transactions paying as fee). With a higher number of pending transactions (i.e., higher network load), not every transaction can be included within the next block of the blockchain which means higher priced transactions get prioritised since block validators earn more per each included transaction. As the gas price is increased, the average latency converges to the average block time (~14-30 seconds) of the blockchain, since higher-priced transactions are almost certainly included within the next block. To evaluate throughput, we submit up to 150 transactions to the test network during a 60 second time frame. Afterwards, we divide the number of confirmed transactions with the time frame of 60 seconds to calculate the transactions per second (TPS). Fig. 2 shows the average transaction throughput of our experiments. The tradeoff that blockchain-based systems make in terms of scalability and security becomes clear with an average throughput of only 1-1.6 TPS. Increasing the gas cost does not seem to improve throughput beyond a certain point. Notably, throughput was measured from a user perspective, i.e., transactions were submitted from a single account. The maximum global throughput of the Ethereum blockchain is roughly 20 TPS.

To sum up, the framework fulfils all functional requirements defined in Section 3. By leveraging blockchain technology, the non-functional requirements of integrity and availability can be fulfilled. Regarding scalability and privacy, the use of blockchain technology poses limitations. Throughput and latency of the framework are constrained, and the inherent transparency of blockchain technology naturally presents a challenge for privacy-sensitive applications.

6 RELATED WORK

As mentioned in Section 1, other works [8, 9] focus on providing generic blockchain-based data provenance solutions for the IoT. However, these approaches remain conceptual

⁶26 July 2019 (<https://ethernodes.org/network/1>)

with no provided implementation. In contrast, the framework presented in the work at hand has been fully implemented and is available as open-source software. Other solutions related to blockchain-based data provenance in the realm of the IoT go beyond conceptual approaches [10–14]. While the approaches presented in [10–12] focus on securing provenance information in supply chains, the works in [13] and [14] focus on a blockchain-based patient monitoring system to provide notifications in case of medical emergencies, and a framework that tracks information about vehicles to resolve disputes between drivers, insurance companies and maintenance server providers in case of accidents, respectively. In general, these solutions demonstrate how provenance concepts can be brought onto the blockchain (e.g., via smart contracts). However in contrast to our work, though IoT technologies are addressed, these solutions are use case-specific and do not provide a generic provenance framework that is required for the heterogeneous nature of the IoT [1, 7].

Provenance in the supply chain has also received attention in the industry. Startups like Provenance⁷ and Everledger⁸ focus on securely tracking goods traveling along a supply chain via blockchain technology. However, their sole focus lies on tracking physical items instead of any kind of data within the IoT. Further, works have also focused on blockchain-based data provenance solutions outside the IoT domain, e.g., blockchain-based data provenance for cloud environments to verify the operation history of data in the cloud [18], or in the scientific field where blockchain technology is used to secure the derivation history of scientific data [19].

As can be seen by the discussion, to the best of our knowledge, there is no other approach which provides a generic framework for tracking data provenance in the IoT.

7 CONCLUSION

The presented framework provides a data provenance solution which is appropriate for the heterogeneous nature of the IoT. By leveraging a generic data model together with a layered architecture of smart contracts, the framework is able to fulfil the functional requirements. By building on blockchain technology, the framework further fulfils the non-functional requirements of integrity and availability but has some limitations regarding scalability and privacy. Multiple current blockchain developments seek to overcome these issues without jeopardising blockchain security, e.g., state channels, zero-knowledge proofs, and sidechains [20]. Thus, in future work, we will evaluate in detail to what extent these approaches provide solutions to the privacy and scalability problem in the context of blockchain-enhanced IoT applications, such as our data provenance framework.

⁷<https://www.provenance.org/>

⁸<https://www.everledger.io/>

ACKNOWLEDGMENTS

The work presented in this paper has received funding from Pantos GmbH within the TAST research project.

REFERENCES

- [1] S. Li et al. “The internet of things: a survey”. In: *Information Systems Frontiers* 17.2 (2015), pp. 243–259.
- [2] M. N. Aman et al. “Secure Data Provenance for the Internet of Things”. In: *3rd ACM International Workshop on IoT Privacy, Trust, and Security*. ACM, 2017, pp. 11–14.
- [3] E. Bertino. “Data trustworthiness—approaches and research challenges”. In: *Data privacy management, autonomous spontaneous security, and security assurance*. Springer, 2014, pp. 17–25.
- [4] J. Freire et al. “Provenance for computational tasks: A survey”. In: *Computing in Science & Engineering* 10.3 (2008), pp. 11–21.
- [5] R. Hasan et al. “Preventing history forgery with secure provenance”. In: *ACM Transactions on Storage* 5.4 (2009). Article no. 12.
- [6] K. Christidis et al. “Blockchains and smart contracts for the internet of things”. In: *IEEE Access* 4 (2016), pp. 2292–2303.
- [7] H. Olufowobi et al. “Data Provenance Model for Internet of Things (IoT) Systems”. In: *Service-Oriented Computing–ICSOC 2016 Workshops: Revised Selected Papers*. Springer, 2017, pp. 85–91.
- [8] G. C. Polyzos et al. “Blockchain-assisted information distribution for the Internet of Things”. In: *2017 IEEE International Conference on Information Reuse and Integration*. IEEE, 2017, pp. 75–78.
- [9] N. Baracaldo et al. “Securing Data Provenance in Internet of Things (IoT) Systems”. In: *Service-Oriented Computing–ICSOC 2016 Workshops: Revised Selected Papers*. Springer, 2017, pp. 92–98.
- [10] K. Toyoda et al. “A Novel Blockchain-Based Product Ownership Management System (POMS) for Anti-Counterfeits in The Post Supply Chain”. In: *IEEE Access* 5 (2017), pp. 17465–17477.
- [11] M. Westerkamp et al. “Blockchain-based Supply Chain Traceability: Token Recipes model Manufacturing Processes”. In: *IEEE International Conference on Blockchain*. IEEE, 2018.
- [12] H. Wu et al. “A distributed ledger for supply chain physical distribution visibility”. In: *Information* 8.4 (2017), 137:1–137:18.
- [13] K. N. Griggs et al. “Healthcare blockchain system using smart contracts for secure automated remote patient monitoring”. In: *Journal of Medical Systems* 42.7 (2018), 130:1–130:7.
- [14] M. Cebe et al. “Block4forensic: An integrated lightweight blockchain framework for forensics applications of connected vehicles”. In: *IEEE Communications Magazine* 56.10 (2018), pp. 50–57.
- [15] Y. L. Simmhan et al. *A survey of data provenance techniques*. Tech. rep. IUB-CS-TR618. Computer Science Department, Indiana University, Bloomington, Indiana, USA, 2005.
- [16] T. Comes et al. “Cold chains, interrupted: The use of technology and information for decisions that keep humanitarian vaccines cool”. In: *Journal of Humanitarian Logistics and Supply Chain Management* 8.1 (2018), pp. 49–69.
- [17] *Cryptographic Tokens*. <https://blockchainhub.net/tokens/>. Accessed 29 May 2019. 2019.
- [18] D. Tosh et al. “Data Provenance in the Cloud: A Blockchain-Based Approach”. In: *IEEE Consumer Electronics Magazine* 8.4 (2019), pp. 38–44.
- [19] A. Ramachandran et al. “SmartProvenance: A Distributed, Blockchain Based Data Provenance System”. In: *8th ACM Conference on Data and Application Security and Privacy*. ACM, 2018, pp. 35–42.
- [20] J. Eberhardt et al. “On or Off the Blockchain? Insights on Off-Chaining Computation and Data”. In: *2017 European Conference on Service-Oriented and Cloud Computing*. Springer, 2017, pp. 3–15.