BlockFlow: Trust in Scientific Provenance Data

Raiane Coelho, Regina Braga, José Maria David, Fernanda Campos, Victor Ströele

Computer Science Post-Graduate Program

Universidade Federal de Juiz de Fora, MG, Brazil.

{raianecoelho, victor.stroele}@ice.ufjf.br; {regina.braga, jose.david, fernanda.campos}@ufjf.edu.br

Abstract. *In scientific collaboration, the data sharing, the exchange of ideas and results is crucial to promote knowledge and accelerate the development of science. Trust is extremely important in this context as well as reproducibility. Although in scientific workflow the provenance has been the basis for reproducibility, in collaborative environments it is necessary to ensure integrity and trustworthiness of this provenance data. One of the technologies that have emerged and can help to address these issues is blockchain. A blockchain-based provenance system for collaborative scientific experiments could lead to a trustworthy environment for scientific experimentation. In this vein, this paper presents the specification of an architecture, named BlockFlow, that provides trust for distributed provenance data.*

1. Introduction

In e-science, there is a demand to support the increasingly collaborative, distributed and multidisciplinary activities, considering that scientific experiments are often conducted through collaborative efforts between different researchers and institutions. In this scenario of scientific collaboration, with the interaction between distributed geographically individuals, data sharing, the exchange of ideas and results is crucial to promote knowledge and accelerate the development of science. However, in this context, where there are several parties involved, trust is extremely important. The lack of trust and transparency in the sharing of information between researchers is a challenge, including the reuse of knowledge acquired in experiments, produced by third parties.

Concurrently, in the scientific community, the reproducibility of experiments is an important issue [Stodden 2010]. In this vein, reports that a disturbing proportion of scientific studies are not reproducible is an increasing concern [Baker 2016]. Faced with this problematic, provenance data in scientific experimentation plays a fundamental role.

Provenance, that describes the origins of a data and the process that derives the data, is a topic that has relevance and has been gaining prominence over the years in several areas and contexts [Herschel et al. 2017]. Although in the scientific workflow the provenance has been the basis for reproducibility, in collaborative environments it is necessary to ensure integrity and trustworthiness of this provenance data. As reported before, on the crisis of reproducibility we do not know whether the provenance was intentionally altered or not. Thus, in the scientific community, the issue of trust is of great importance.

There is a need to guarantee the collection and unchanging storage of this provenance data [Liang et al. 2017]. One of the technologies that have emerged recently and can be key to address the issues of reproducibility, and trust in provenance data, is blockchain [Nakamoto 2008]. The blockchain is specified for collaborative and distributed scenarios since it has been considered a promising technology that can improve the way we interact.

On the other hand, in this scenario of collaborative and distributed environment, SECOs (Software Ecosystems) [Manikas 2016] has emerged as a paradigm to understand the dynamics and heterogeneity in the development of collaborative software. Based on this ecosystem approach, [Freitas et al. 2015] specified the E-SECO (E-Science Software ECOsystem) platform. This platform can manage and support all stages of scientific experiment lifecycle. A blockchain-based provenance system for collaborative scientific experiments could lead to a trustworthy environment for scientific experimentation, allowing a transparent audit trail of all data that is collected, processed and accessed by different workflows, geographically distributed.

Considering blockchain a disruptive, distributed and immutable ledger, and mainly by focusing on data, the main contribution of this work is the specification of an architecture, named BlockFlow. This architecture is part of the E-SECO platform to provide trustworthy to provenance data. Our proposal is a blockchain-based architecture for storing provenance data.

Although in the literature, there are approaches that deal with the use of blockchain for provenance data [Liang et al. 2017], [Ramachandran et al. 2018], these approaches do not offer guidelines that can support the management of provenance data, in the collaborative scenario in a scientific experimentation platform.

This paper has 5 sections, including this introduction. Section 2 discusses some basic concepts related to our approach. Section 3 presents BlockFlow and some related works. Section 4 presents a BlockFlow implementation and finally, section 5 concludes the paper.

2. Background

The blockchain is a technology proposed by [Nakamoto 2008], to solve the problem of double spending in virtual currencies. A blockchain can be defined as a distributed ledger, which offers an environment of immutability, security, with greater reliability, transparency, privacy, and efficiency as well as can be used for provenance data related applications [Liang et al. 2017]. Users can send transactions between two or more parties of mutually untrusting, without the need for a single centralized authority [Nakamoto 2008]. Fundamentally, blockchain consists of blocks, containing a set of transactions. Each block has a timestamp associated and a link to a previous block [Nakamoto 2008] [Vukolić 2015]. This operation establishes a link between the blocks created, thus creating a chain of blocks or blockchain.

There are two different blockchain types, i.e., permissionless or public ledgers and permissioned or private ledgers. In permissionless blockchain, anyone can join and leave in the ledger, as well as read, write and validate transaction, without the need for a central authority. Bitcoin [Nakamoto 2008] and Ethereum[1] are instances of permissionless

[1] https://www.ethereum.org/

blockchains. In permissioned blockchain, the participation is permissioned and the access is restricted to a certain number of participants, which are known to each other. Here, a central entity decides who read and write transaction in the blockchain. Hyperledger Fabric[2] and R3 Corda[3] are instances of permissioned blockchains. In permissioned, or private ledgers, often utilizes the Byzantine Fault Tolerance (BFT) [Vukolić 2015] as its consensus algorithm.

The blockchain initiative can be attractive to scientific experimentation, considering that collaborative and distributed experimentation involves the production of provenance data and the sharing of this information needs trust. Blockchain implementation can provide this trust. However, transactions in permissionless blockchain are validated by miners through economic incentives, which can incur in high costs for transactions. In this way, opted to use a permissioned blockchain, i.e., Hyperledger Fabric or just Fabric, which is an open-source [Androulaki et al. 2018] blockchain platform, where all participants have known identities, is an ideal scenario for privacy and confidentiality of provenance data. The decentralization and security characteristics of blockchain have attracted the use of smart contracts for various applications. A smart contract is a self-executing code that verifies pre-defined terms and conditions. In our approach, using the Hyperledger Fabric, a smart contract is called *chaincode*, and executes distributed applications written in common programming languages (e.g., Go, Java, Node.js).

In general, scientific distributed experiments are modeled in different Scientific Workflow Management Systems (SWfMSs), such as Kepler[4], Taverna[5], among others. These SWfMSs automatically capture provenance data. However, in general, their proprietary models make it difficult to share information, which could hinder collaborative research. Therefore, in order to facilitate data provenance´s capture and integration, models such as PROV [Groth and Moreau 2013] emerged. For specific domains, there are model extensions, as is the case of scientific workflows. In this sense, ProvONE [Cuevas-Vicenttín et al. 2015] was specified, which details scientific processes, ports, and data links.

3. BlockFlow

The process of scientific experimentation involves interactions between researchers and geographically distributed institutions. In this context, [Freitas et al. 2015], specified the E-SECO (E-Science ECOsystem) platform, based on the SECO approach [Manikas 2016]. However, E-SECO lacks a system that provides trustworthy for distributed scientific experimentation. In this way, in order to support a trustworthy environment in the context of E-SECO platform, BlockFlow architecture was specified. BlockFlow is an architecture that uses blockchain to provide trustworthy and immutable provenance in the context of E-SECO.

---

[2] https://www.hyperledger.org/projects/fabric

[3] https://www.corda.net/

[4] https://kepler-project.org/

[5] http://www.taverna.org.uk

In order to specify BlockFlow, the following requirements were considered: (i) It must be able to capture prospective, retrospective and evolution [Missier et al. 2013] provenance data, allowing the interoperability of these data that are generated by different SWfMSs, in their proprietary models; (ii) The architecture should provide the storage of provenance data collected in an immutable and trustworthy way; (iii) Allows the audit and trail of provenance data.

As stated before, in general, collaborative scientific workflow provenance, are modeled in proprietary formats used by different SWfMSs. Therefore, ProvONE model [Cuevas-Vicenttín et al. 2015] is used as the integration model. In this scenario, trust in provenance data, obtained from the collaborative scientific workflows, is also crucial. The objective is to support the reproducibility of scientific results. For this, BlockFlow proposes the storage of provenance data, in the form a block in the blockchain so that provenance data is authentic and tamper-proof.

Figure 1 presents the BlockFlow architecture, with four main layers: **Collaborative scientific workflow**, in this layer, collaborative SWfMSs are used, by multiple scientists that can be geographically distributed, designing, executing monitoring, validating and tracking provenance collaboratively; **Wrapper,** that capture and translates provenance data from the SWfMSs to ProvONE model, it is possible to capture prospective, retrospective and evolution provenance and make data provenance integration; **Blockchain network,** all the provenance data will be stored in blocks, assuring a trustworthy and immutable provenance data and these data can be shared and accessed by scientists; **View**, through this layer, scientists can interact with the application and provenance data can be monitored and audited. Figure 2 usage scenario was detailed in subsection 3.1 a better understanding the BlockFlow architecture.

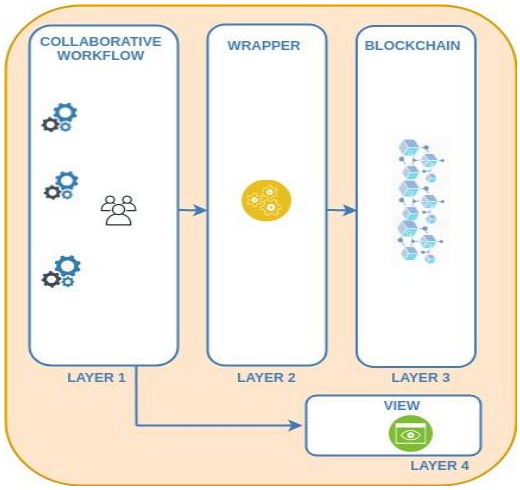

Figure 1. Architecture Overview

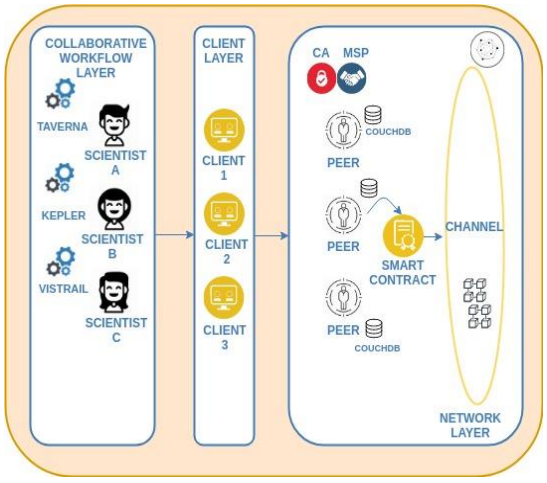

Figure 2. Usage Scenario

In the literature, there are various authors have that deal with the use of blockchain for trustworthy provenance data. [Liang et al. 2017] proposed ProvChain, a blockchain-base data provenance architecture, to provide tamper-proof provenance for cloud storage applications. [Ramachandran et al. 2018] proposed the decentralized system SmartProvenance that uses blockchain to store provenance data. [Pahl et al. 2018] proposed a blockchain-based architecture for trustworthy processing of IoT edge architectures. Although these approaches deal with provenance stored using blockchain

technology, no one provides trust to scientific experimentation, dealing with distributed SWfMSs and also providing a secure private channel to scientists collaborating in an experiment. Besides, these works do not provide a query layer to monitor and audit provenance in scientific experimentation.

3.1 Usage Scenario

In order to a better understanding of BlockFlow, a usage scenario was detailed. Let us consider the scenario presented in Figure 2. The goal is to support the reproducibility of scientific results.

In this scenario, multiple scientists are geographically distributed to collaboratively design, execute, monitor, validate, track provenance, and manage scientific experiments. Let us suppose that scientist A uses Taverna, and scientist B uses Kepler, scientist C uses Vistrails[6] as SWfMSs, as they participate in an experiment, they need to capture and share provenance for task collaboratives. BlockFlow uses Fabric [Androulaki et al. 2018] as the Blockchain platform. A Fabric blockchain consists of a set of nodes that form a network. Each node maintains the state of the ledger and log of transitions through Apache CouchdDB[7] or LevelDB[8]. The nodes represent scientists that belong to the experiment. These scientists can collaborate and can store provenance data captured (during workflow design or execution) in blockchain fabric. For this, BlockFlow specifies a channel. Channel is a private blockchain overlay on the network, to allow data isolation and confidentiality [Androulaki et al. 2018]. In BlockFlow, there is a channel for each experiment. To transact in the channel, i.e., store or query provenance data, all nodes that participate in the network need to have an identity. Identities in Fabric is provided by CA (Certificate Authority) [Androulaki et al. 2018], that together with the MSP (membership service provider) are responsible for membership enrollment by issuing enrollment certificates and transaction. In channels, we implement an access control list, where different access type can be specified. This configuration specifies who can query and update operations for *chaincode* execution in the channel. In our case, they are the scientists that are part of the experiment. Therefore, Figure 2 illustrates the scientists in the layer "collaborative workflow", and using different SWfMSs to execute their workflows, that is part of an experiment. Each workflow provides provenance data that are captured and sent to the blockchain. For this, the scientists connect to the Fabric client in "client" layer (Figure 2), that communicates with the peer in "blockchain network" layer and can invoke and execute the transaction through *chaincode* in the channel that it belongs, and then record the provenance data. In this way, the scientists can audit all provenance data that is collected, processed and accessed by different scientists that participate in the experiment.

4. BlockFlow in Action

This section details how BlockFlow can support the researcher in the process of scientific experimentation. In this study, we consider the extensions provided by [Freitas et al. 2015] and [Sirqueira et al. 2016], including the mechanisms that can enhance

---

[6] https://www.vistrails.org

[7] http://couchdb.apache.org/

[8] http://leveldb.org/

collaboration among scientists provided by E-SECO. In addition to these extensions, BlockFlow provides trust to provenance management.

Therefore, in order to present this feasibility study, a blockchain network environment was specified so that the nodes of E-SECO could share trustworthy provenance data. Our study uses a workflow available at myExperiment repository http://www.myexperiment.org/workflows/2258.html. This workflow performs a search for sequences similarity through the BLAST (Basic Local Alignment Search Tool) algorithms, from protein sequences in the FAST format as input, using NCBI (National Center for Biotechnology Information) blast of the EBI (European Bioinformatics Institute) services. This workflow consists of two workflows BI_NCBI_BLAST and Fasta_string_to_fasta_list. In our scenario, the experiment is executed using two different SWfMSs. Figure 3 (a) shows the first part of the workflow specification in the Kepler SWfMSs, whereas Figure 3 (b) presents the specification of the second part in Taverna SWfMSs.

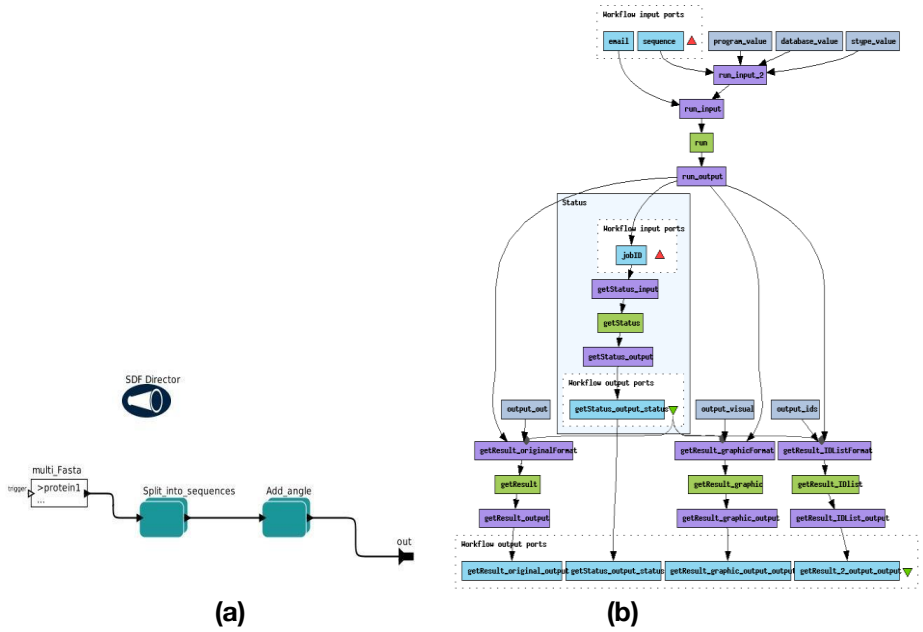

Figure 3. (a) Kepler, (b) Taverna.

In this collaborative scenario, the experiment was designed by two researchers, geographically distributed. BlockFlow was used considering the Fabric version 1.1, in the Docker[9] container environment. In this study, the network was specified using two nodes. One real machine, running Taverna SWfMSs, and one virtual machine, running Kepler SWfMSs. These machines are the peers in the blockchain network and belong to a channel, to allow data isolation and confidentiality.

The capture of the provenance data is provided by the Wrapper layer in BlockFlow, as shown in Figure 1, using a web service, implemented in Node.js that communicates directly with Taverna and Kepler and capture provenance data in real-time. After translate provenance data from the SWfMSs to provONE model, BlockFlow uses smart contracts for accessing shared ledger and store provenance data. All the provenance

[9] https://www.docker.com/

data are stored in blocks. The detailed how the provenance data are collected in from different SWfMSs and translated for the provONE model is not in the scope of this work. Figure 4 shows an example of all transaction and provenance data in BlockFlow architecture.

Figure 4. (a) All transaction in BlockFlow, (b) Provenance data.

Using BlockFlow, we are able to express queries over provenance data. Table 1 presents some queries executed using BlockFlow and Apache CouchDB.

Table 1. Queries

#	Queries specification	Queries
Q1	Retrieve all programs with their input and output ports for each workflow executed the of experiment.	{"selector": {"docType": {"$eq": "program" }, "idWorkflow" :{ "$eq": "idWorkflow"},}, "fields": ["hasInPort","hasOutPort", "nameProgram", "created"]}
Q2	Retrieve all programs executed by the experiment.	{ "selector": { "docType": { "$eq":programExecution" }},"fields": ["idProgramExecution", "programName", "startTime","endTime"]}

5. Conclusions

In this paper, we presented the BlockFlow architecture, a blockchain based data provenance system for collaborative scientific experiments. The BlockFlow provides transparency and trustworthiness for collaborative scientific experiments, storing data that is collected, processed and accessed by different SWfMSs. A usage scenario was detailed, and some related works were discussed.

As future work, we intend to implement the view layer. In this way, we aim to facilitate the interpretation of the data by researchers. In addition, we intend to carry out experimental studies to evaluate the approach, considering different types of scientific experiments, reproduction and reuse contexts.

References

Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., & Muralidharan, S. (2018, April). Hyperledger fabric: a distributed operating system for permissioned blockchains. In Proceedings of the Thirteenth EuroSys Conference (p. 30). ACM.

Baker, M. (2016). 1500 scientists lift the lid on reproducibility. Nature News, 533(7604), 452.

Cuevas-Vicenttín, V., Ludäscher, B., Missier, P., Belhajjame, K., Chirigati, F., Wei, Y., and Altintas, I. 2015. ProvONE: A PROV Extension Data Model for Scientific Workflow Provenance.

Freitas, V., David, J. M., Braga, R., and Campos, F. (2015). An architecture for scientific software ecosystem. In 9th Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems (WDES 2015), pages 41–48. (in Portuguese)

Groth, P., and Moreau, L. 2013. PROV-Overview. An overview of the PROV Family of Documents.

Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from?. The VLDB Journal—The International Journal on Very Large Data Bases, 26(6), 881-906.

Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., & Njilla, L. (2017, May). Provchain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (pp. 468-477). IEEE Press

Manikas, K. (2016). Revisiting software ecosystems research: A longitudinal literature study. Journal of Systems and Software, 117:84–103.

Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttín, V., & Ludäscher, B. (2013). D-PROV: Extending the {PROV} Provenance Model with Workflow Structure. In 5th {USENIX} Workshop on the Theory and Practice of Provenance (TaPP 13).

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Pahl, C., El Ioini, N., Helmer, S., & Lee, B. (2018, April). An architecture pattern for trusted orchestration in IoT edge clouds. In Fog and Mobile Edge Computing (FMEC), 2018 Third International Conference on (pp. 63-70). IEEE.

Ramachandran, A., & Kantarcioglu, M. (2018, March). SmartProvenance: A Distributed, Blockchain Based DataProvenance System. In Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy (pp. 35-42). ACM

Sirqueira, T. F., Dalpra, H. L., Braga, R., Araujo, M. A., David, J. M. N., and Campos, F. (2016). E-SECO proversion: Manutenção e evolução de experimentos científicos. In BreSci – 10º Brazilian e-Science Workshop, pages 253–260. CSBC.

Stodden, V. (2010). The scientific method in practice: Reproducibility in the computational sciences.

Vukolić, M. (2015, October). The quest for scalable blockchain fabric: Proof-of-work vs. BFT replication. In International workshop on open problems in network security (pp. 112-125). Springer, Cham.