Integrating Blockchain for Data Sharing and Collaboration Support in Scientific Ecosystem Platform

Raiane Coelho
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
raianecoelho@ice.ufjf.br

Regina Braga
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
regina.braga@ufjf.edu.br

José Maria N. David
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
jose.david@ufjf.edu.br

Mario Dantas
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
mario.dantas@ufjf.edu.br

Victor Ströele
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
victor.strole@ice.ufjf.br

Fernanda Campos
Department of Computer Science
Federal University of Juiz de Fora
Juiz de Fora, Brazil
fernanda.campos@ufjf.edu.br

Abstract

Nowadays, scientific experiments are conducted collaboratively. In collaborative scientific experiments, we must consider aspects such as interoperability, privacy, and trust in shared data to allow the reproducibility of the results. A critical aspect associated with a scientific process is its provenance information, which can be defined as the origin or lineage of the data that helps understand the scientific experiment results. Another concern when conducting collaborative experiments is confidentiality, considering that only authorized personnel can share or view results. In this paper, we propose BlockFlow, a blockchain-based architecture, to bring reliability to the collaborative research, considering the capture, storage, and analysis of provenance data related to a scientific ecosystem platform (E-SECO).

1. Introduction

In the scientific community, collaboration and data sharing among researchers are essential to support scientific advances [1]. Researchers are encouraged to share resources, opinions, and conduct scientific experiments among geographically distributed groups. However, we must also consider several challenges in collaborative scientific experiments, such as reproducibility, privacy, transparency, and interoperability.

The reproducibility of scientific experiments is critical and an important issue [2]. However, a considerable amount of scientific research loses its credibility because they are non-reproducible [2], and they do not have a specific mechanism to control the steps and historical data related to the experiment. We built every new scientific discovery through an iterative process, based on existing knowledge, so if we cannot reproduce existing knowledge, we are wasting a lot of effort, resources, and time. For the reproducibility of scientific experiments, provenance data [3] plays a key role. Provenance or data lineage is metadata, which describes the origin of data and the processes and transformations that originate it. For scientific experiments, provenance is considered essential to support both the reuse of computational experiments, the interpretation of results, and the diagnosis of problems [4]. In this sense, provenance data on which scientific findings are based must be reliable [5].

Another concern when conducting collaborative experiments is the confidentiality of provenance data, considering that only properly authorized personnel can share or view results. Transparency is another critical issue to guarantee that researchers will have confidence in the collaborative experiment's conduction. Lastly, provenance data integration is critical, considering that researchers use heterogeneous scientific software [6] to execute their experiments. Lacking provenance data integration support makes it difficult to share heterogeneous information, hindering the sharing of knowledge.

In this way, it is not enough to share data or activities among researchers' groups to collaborate. It is essential to ensure scientific reproducibility and correct interpretation of scientific data among geographically distributed researchers.

Complex experiments involve interactions between geographically distributed researchers. We must consider aspects such as the use of large amounts of data and the need to be supported by distributed computing resources and services. Besides, experiments require intense relationships among resources and applications that support the scientific workflow. In this context,

URI: https://hdl.handle.net/10125/70642
978-0-9981331-4-0

scientific institutions had to open their frontiers to collaborate with external partners, arising a new concept of scientific development. This concept encompasses several software solutions, scientific institutions, and scientific software developers that can adhere to a shared Scientific Software Ecosystem (SSECO) platform.

In this vein, to support collaboration and interaction between geographically distributed scientific partners, the E-SECO (E-Science Software Ecosystem) platform was specified [7]. The E-SECO platform manages all stages of the life cycle of collaborative scientific experimentation and the capture of provenance data, through the support of a peer-to-peer network. Each node of the network has an E-SECO data repository, storing data in a decentralized manner. However, although E-SECO's data repository is decentralized and shared among its users, it does not have a mechanism that provides trust both for shared provenance data and for the scientific collaboration process.

In this sense, the blockchain [8] paradigm has emerged, proposing decentralization, collaboration, disintermediation, and a sense of trust. Blockchain-based systems smooth the path to collaborative and distributed scientific organizations to build mutual trust.

We argue that a promising approach can be to use blockchain mechanisms and provenance data to smooth the scientific collaboration process.

Considering the E-SECO platform, this work's main contribution is the specification of a blockchain-based architecture, aiming to bring reliability to the collaborative research in the E-SECO platform. Also, there is an effort to provide privacy, reproducibility, transparency, and interoperability in data share. These requirements are essential to establish trust for data in the research method and obtained results.

This article is organized as follows. Section 2 presents the background. Section 3 describes the proposed solution and E-SECO platform. Section 4 presents a feasibility study that aims to assess the implemented solution. Section 5 discusses related work. Section 6 presents the final considerations, as well as future work.

2. Background

2.1. Data provenance in Collaborative Research

Considering the current scientific research scenario, experiments are guided and executed through Workflow Management Systems (WfMS). In this sense, a critical aspect associated with a scientific process is its provenance information, which can be defined as the origin or lineage of the data that helps to understand the results of scientific experiments [3]. Provenance or data lineage is metadata, which describes the origin of data and the processes and transformations that originate it. More formally, provenance is the metadata that describes entities, data, processes, activities, and people involved in the process of creating a product [9].

In collaborative scientific environments, provenance helps scientists to interpret results and diagnose problems along the experimentation process. However, different scientists can execute part of the experiment on different WfMS, such as Kepler[1], Taverna[2], among others. Some of these WfMS automatically capture provenance data, mainly using a standard model, such as PROV [10], but in some cases with some proprietary extensions. Others use only their proprietary models. Therefore, this provenance data heterogeneity makes it difficult to interpret, share, and combine the information. In the context of heterogeneous data, to promote provenance interoperability and sharing, several models were implemented, such as the OPM [9] and PROV [10], recommended by the W3C. OPM was widely adopted by the scientific community but was discontinued and replaced by the PROV model. PROV renamed some entities and relationships and added new relationships to express provenance in a more general sense. There are different forms to support and capture provenance information. Lim et al. 2010 [11] discuss two types of provenance: prospective and retrospective. According to Freire et al. [12], there is also a third type called evolution provenance.

- Prospective: captures the structure and static context of a workflow. It expresses the steps (or recipe) to be followed to generate a data set.
- Retrospective: it is associated with information about the execution of a workflow, that is, information about the steps taken to derive a data set. More specifically, it is a detailed log of the execution of each task in a workflow.
- Evolution**:** reflects the changes made between two executed versions of the workflow. Records the history of the workflow evolution, keeping all changes applied throughout its life cycle. This provenance type is important in BlockFlow considering that when provenance is stored it cannot be changed. In this sense, when a workflow task is modified, BlockFlow deals with evolution provenance.

The PROV model is generic and can be extended to capture more properly the provenance in specific

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

[2] https://taverna.incubator.apache.org/

domains. A PROV extension adapted to the context of scientific workflows, called ProvONE [13], was developed. ProvONE [13] can capture prospective, retrospective, and evolution provenance and has specific entities and relationships for the scientific workflow and data representation process.

2.2. Blockchain

Blockchain is an immutable ledger, shared, decentralized, that maintains a sequence of chronological, encrypted, and synchronized blocks, over a peer-to-peer network [8]. Each block is 'chained' to the previous block by including the block's hash value. These blocks contain a list of transactions that occur between the participating peer nodes of the network. New blocks are added to the end of the chain and existing transactions cannot be updated or deleted (thus, blockchain provides immutable data storage). To add new blocks to the network, the participants' peers must validate the block. This validation is done through consensus mechanisms, ensuring that data is decentralized among several nodes that hold identical information and that no single actor holds the network's complete authority. There are currently several consensus mechanisms for blockchain, such as proof-of-work, proof-of-stake, and byzantine fault tolerance [14].

Blockchain networks can be classified as permissionless or permissioned. This classification determines who can participate or transact on the network and determine the identity of its participants. In permissionless networks such as Bitcoin[3] and Ethereum[4] anyone can join, transact, leave the network, or verify any transaction. The privacy or confidentiality of the participating is maintained using public-key cryptography. However, the transaction data is not private, being necessary to keep the data privacy, cryptographic means. A permissionless blockchain network generally uses the proof-of-work or proof-of-stake consensus mechanisms to prevent fraudulent transactions. In permissioned networks such as Hyperledger Fabric [15] and Corda[5], a group of known nodes controls the network, and only authorized nodes can participate. Blockchain permissionless can have positive effects on collaborative scientific workflow processes. However, due to data privacy and intellectual property concerns, blockchain with permissioned becomes a more realistic option for collaborative scientific workflows with provenance data sharing. PoW and PoS are two of the most common consensus techniques in permissionless blockchains that guarantee trust, however the mining process is time consuming. In contrast, permissioned blockchain leverages faster protocols to achieve consensus [14]. In its early beginning, blockchain was only seen as the technology behind most existing cryptocurrencies. However, this technology is not limited to these applications and is used today in different contexts. The scientific community can benefit from the blockchain technology in order to provide trust in a collaborative environment. Decentralization, transparency, immutability, and trust are features that the scientific community can take advantage of blockchain technology. Therefore, the proposal described in this article aims to provide a collaborative and reliable environment, supported by blockchain technology, for scientific experiments. In order to provide this collaborative environment, the blockchain technology chosen was Hyperledger Fabric [15]. Hyperledger Fabric is an open-source project maintained by the Linux Foundation. Network access is restricted to authorized persons, i.e., usually composed only by people who have a common interest, which can establish greater trust in the network. The Hyperledger Fabric network consists of a set of geographically distributed peer nodes that maintains the state of the ledger and the log of transactions through Apache CouchdDB[6] or LevelDB[7]. There are different types of peer nodes in the Hyperledger Fabric. The ordering peer is responsible for receiving customer transactions and specifying how these transactions will be stored. It uses Apache Kafka[8], which allows a distributed storage with fault tolerance. The network consensus mechanism uses the Zookeeper technology, which applies a version of Paxos consensus mechanisms [16]. Their transactions are controlled and generated through smart contracts (chaincodes). Chaincode is a software that reads and updates the ledger state. This software can be written in programming languages, such as Go, Java, and Node.js. Nodes peers communicate using channels. The channels maintain privacy, confidentiality, and isolate activities between authorized parties. In addition to transacting, the participating nodes need to enroll and have identities. Identity records are provided by the Certificate Authority (CA), which also issues certificates to be used to sign transactions. There is another essential component for identifying nodes with CA: the Membership Service Provider (MSP) responsible for mapping certificates between nodes.

[3] https://www.bitcoin.com/
[4] https://ethereum.org
[5] https://www.corda.net/
[6] https://couchdb.apache.org/
[7] https://dbdb.io/db/leveldb
[8] https://kafka.apache.org/

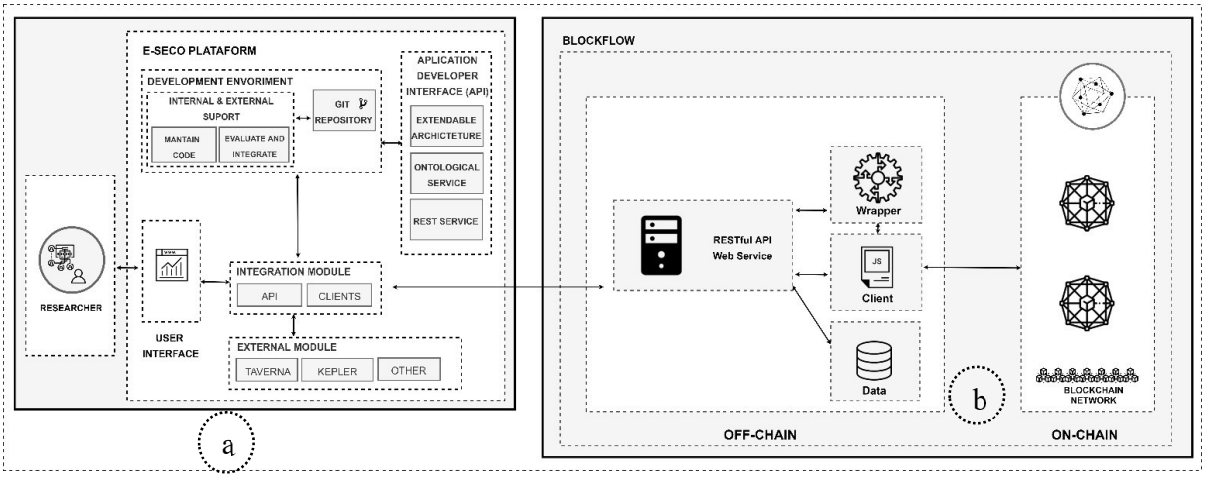

Figure 1. Overview of E-SECO platform (a) and BlockFlow architecture

3. BlockFlow Architecture

3.1. E-SECO platform

In the modern science scenario, an eScience infrastructure needs to provide an environment capable of addressing heterogeneous data production, reproducibility, and providing a collaborative and reliable environment for distributed groups of researchers. Therefore, to support the collaboration between geographically distributed scientists, the E-SECO platform was proposed [7]. This platform is based on the Software Ecosystem [17], focused on the eScience domain. The E-SECO platform manages the entire life cycle of a scientific experiment, also considering provenance data management, with the support of a peer-to-peer network. As shown in Figure 1a, this platform comprises a development environment, an integration layer, among other layers.

An experimentation process usually traces the following steps in E-SECO. During the problem investigation step, scientists look for similar experiments, interact with other researchers using the E-SECO platform, define their goals, and break down the experiment into smaller steps. In the experiment prototyping step, scientists build a prototype by designing workflows and reusing available assets. Scientists access artifacts persisted in E-SECO related repositories. Therefore, researchers can explore the assets and reuse their components to produce new ones and contribute with new artifacts during the experiment prototype step. As a final step, researchers analyze and publish their results and contributions, using the collaboration support provided by E-SECO.

E-SECO enables the storage of information about the experiment process in a detailed way, including experiment steps, execution conditions, input and output data, iterations, results analysis, guaranteeing experiment quality. The Provenance and Context Layer is responsible for capturing, storing, and sharing information from scientific experiments. Although the E-SECO data repository is decentralized, it does not have a system that provides trust for data storage and sharing, which is essential for collaborative research considering geographically distributed scientists. In order to cope with this problem, we specified BlockFlow architecture. In the next sections, the BlockFlow architecture is detailed, considering the E-SECO platform.

3.2. Architecture overview

As stated before, BlockFlow architecture was specified to bring trust for collaborative distributed experimentation in the E-SECO platform. We considered the following requirements when designing the architecture: (i) Network: scientists must be able to create Blockchains networks, allowing that peers that represent scientists in the E-SECO platform can have a trusted environment. (ii) Reproducibility: provenance data must be collected and stored, immutably, and trustworthy. Trust in provenance data obtained from collaborative research is crucial to support the reproducibility of scientific results. (iii) Provenance data sharing transparency: Blockchains are fundamentally transparent, where data and interactions are visible to all participants in the blockchain network. (iv) Privacy: provenance data should be shared between authorized personnel. Blockchain Hyperledger Fabric implements solutions to guarantee that private data can be shared only between authorized scientists. (v) Interoperability: provenance data collected in heterogeneous scientific applications should be integrated. Researchers use heterogeneous scientific

applications in WfMS, and the data collected in these heterogeneous applications and workflows should be integrated, enabling the analyzes and comparisons to derive conclusions. Provenance data capture, using a standard model, such as ProvONE, can be used in this context. We developed BlockFlow as an architecture that can connect with other applications using an API. Figure 1 b presents an overview of the architecture. The main components of BlockFlow are discussed below:

3.3. RESTful WebService API Layer

The RESTful web service API layer allows that BlockFlow can be integrated with any other platform or application, based on communication via REST web services and HTTP. Its main objective is that platforms and applications users (scientists) can easily create blockchain networks to collaborate, ensuring trust and reproducibility for scientific experiments. Specifically, through this layer, scientists can request BlockFlow to i) create blockchain networks for an experiment; ii) store and query provenance data. iii) accomplish operations related to Blockchains networks, such as install chaincode, instantiate chaincode, join peer in the channel, and configure blockchain network, among others.

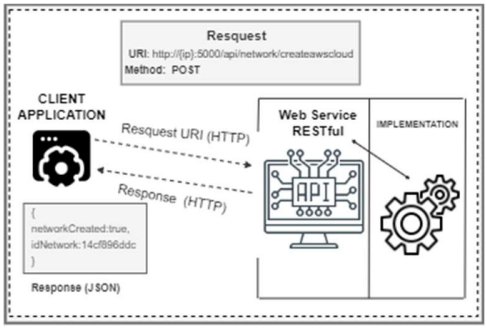

Figure. 2. An example of a request to RESTful WebService API.

Figure 2 presents an example of the request-response flow between the E-SECO and the BlockFlow RESTful layer. This request asks the BlockFlow to create a blockchain network so that researchers can collaborate on their experiment. This layer's main advantage is that it allows one software to communicate with another without knowing the implementation details.

3.4. Client Layer

The Client Layer allows applications to connect to the blockchain network and the peers. This layer is composed of a set of methods to interact with the ledger. Figure 3 presents a flow of requests for this layer. An application can request the RESTful WebService API to send or query a transaction for the blockchain network. The RESTful web service API will request the Client Layer to connect to a peer in the blockchain network. The peer then invokes/query the chaincode to send a provenance data transaction.

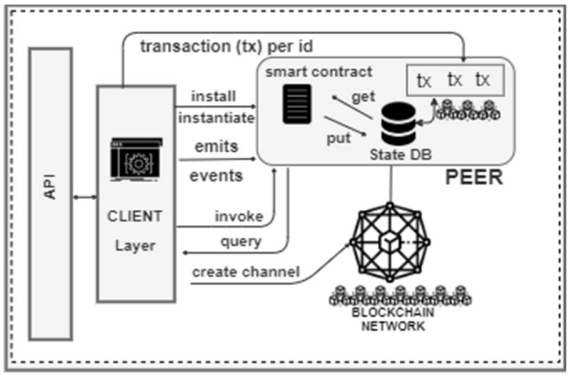

Figure. 3. Client Layer and request Peer.

3.5. Wrapper Layer

Considering that scientists in collaborative experiments can perform part of the experiment in heterogeneous environments, through different WfMS, this layer translates the provenance data to the ProvONE model.

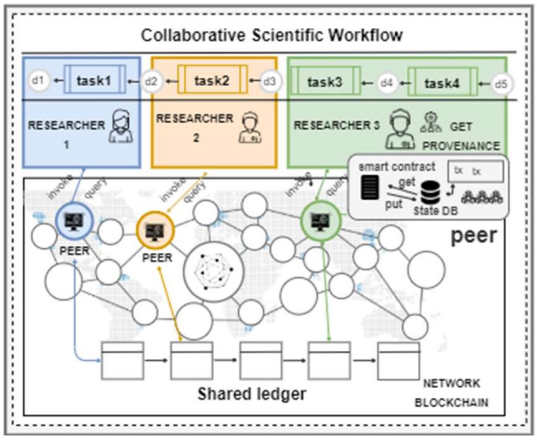

Figure. 4. Collaborative Scientific Workflow in Blockchain Network.

Considering the workflow shown in Figure 4, task T1 can be performed by researcher 1 at location L1, represented by the color blue, while task T2 by researcher 2, represented by the color yellow, or even task T3 and T4 can be performed by researcher 03 represented by the color green.

The mapping of provenance to the ProvONE model (Figure 5) occurs by observing the invocation of

tasks and the life cycle of the data sets consumed or produced during workflow execution. To collect provenance data from the distributed peers, represented by scientists collaborating in an experiment, BlockFlow provides a web service component. The capture is in real-time and WfMS' independent. The web service component is connected to each workflow task, and then it captures the provenance data. In the header of each web service, the user's token received when authenticating must be sent to the system. This procedure is a way of confirming authenticity and retrieving user identity as a node belonging to the blockchain network and ensuring that data is later signed and transacted on the blockchain network.

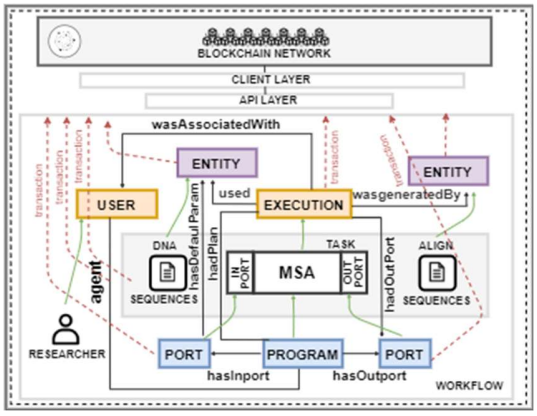

Figure. 5. Tasks' mapping to the ProvONE model.

Figure 5 shows an example of a provenance task mapping to the ProvONE model through the wrapper layer. The workflow task presented in Figure 5 consists of a Multiple Sequence Alignment (AMS) activity, which receives as input a file containing DNA sequences and generates the alignment of that sequence as an output. Considering the ProvONE model, its classes and relationships (Figure 5), the task (AMS) is mapped to the *Program* class of the ProvONE model, and its input and output ports to the *Port* class, where the *hasInport* and *hasOutport* relationship relate them, respectively, to a *Program* (the green arrows express the correspondence between the elements of the workflow and the classes of the ProvONE model).

The task, when executed, is mapped to the *Execution* class, and the DNA SEQUENCES input file is mapped as an *Entity* (we store the hash and the path of the input and output data, instead of the real data, to reduce the volume of provenance data), which has the *hasDefaultParam* relationship with an Input Port and which is used by an Execution. The DNA ALIGN file generated as an output *wasGenerationBy* the execution of the MSA task. The execution, related to the *Execution* class, of the task (MSA) *wasAssociatedWith* a researcher, mapped to the *User* class, which relates to *hadPlan* with an instance of the *Program* class. The Wrapper layer, after these mapping, sends the collected provenance information (classes and relationships) to the "Client Layer" which then connects to a blockchain network peer, as shown in Figure 5, which makes calls to the chaincode. The chaincode then stores the provenance data in the state DB and in the blockchain file system, which is then shared between the blockchain network nodes.

3.6. Data Layer

This layer allows the storage of BlockFlow configuration data. Figure 6 (a) shows the user interface for creating a blockchain network environment in BlockFlow. The configuration data, such as the name of the experiment, organizations, and the number of their peers, are stored in the data layer. Therefore, considering the complexity of a wide range of blockchain-based applications and services, the specification of blockchain network environments at BlockFlow is transparent to the user, with the help of the information persisted in the data layer. The entire configuration is entirely autonomous, and scientists do not have to worry about the configurations between the different components of the blockchain technology used.

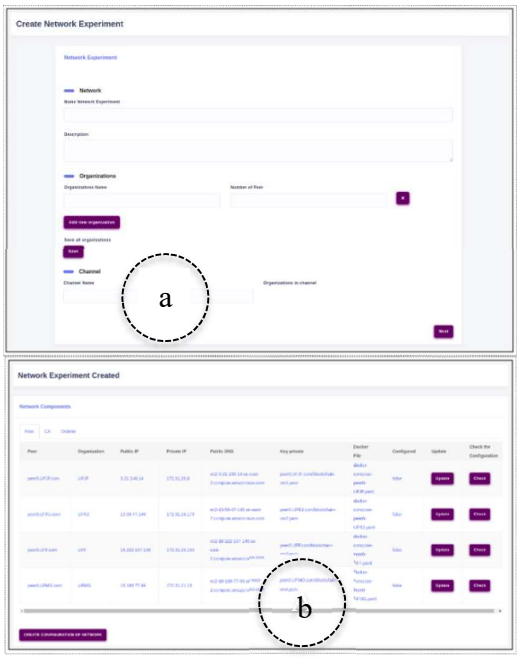

Figure. 6. The user interfaces for blockchain network creation (a) and Hyperledger Fabric components (b).

3.7. Blockchain Network

The blockchain network consists of the geographically distributed peers that represent the researchers collaborating in an experiment. Figure 4 represents a collaborative workflow composed of scientists who can be in geographically distributed locations. All provenance data collected by them will be stored in blocks, enabling immutable storage. This stored data will be shared among all, thus ensuring transparent access to provenance data, collected, and processed by different peers, geographically distributed.

A blockchain-based provenance system for collaborative scientific experiments can lead to a reliable scientific experimentation environment since the collected provenance cannot be manipulated without leaving a trail. This layer guarantees transparency, immutability, and reliability for provenance data of scientific experiments.

3.8. Implementation details

The architecture was divided into two modules on-chain and off-chain. The RESTful API web service, the Client, Wrapper, and Data layers compose the off-chain module. The RESTful API web service and Wrapper Layer was implemented using the Node.js[9] technology. The Client layer was implemented using the Hyperledger Fabric SDK[10] for Node.js in order to interact with the Blockchain Network. The Data layer was implemented using MongoDB[11] database.

The on-chain module was implemented using the Hyperledger Fabric platform. Each component, such peers, CAS, CouchDB, among others, are docker[12] and were specified in yaml[13] files, which need to be initialized using the docker compose[14]. To store and retrieve information in the blockchain, it is necessary to use chaincodes implemented using the go[15] programming language.

4. Feasibility study

A feasibility study is an assessment of the practicality of a proposed project or system. It determines whether the solution considered to accomplish the requirements is practical and workable in the software. Therefore, this feasibility study's main objective was to evaluate the operation of the architecture's technological components.

We used BlockFlow implementation to specify a scientific experimentation trustful environment using E-SECO, to allow the conduction of a collaborative scientific experiment so that geographically distributed researchers can share and integrate their provenance data and scientists can compare and analyze methods and results in a geographically distributed trustful environment.

4.1. Collaborative scenario

The development of new drugs is not a trivial task. Typically, the time to develop a candidate drug is about 10-15 years [18]. Thus, the discovery of new drugs can be drastically accelerated through computational tools such *as in-silico* workflows [19]. In [19], the authors proposed a scientific workflow called SciPPGx, which aims to design candidate drugs. Considering the importance of research related to the development of new drugs and the importance of ensuring the reproducibility of their findings [18] and the importance of collaboration between scientist with different expertise, this scientific workflow was used to verify the feasibility of using BlockFlow in a collaborative environment and their ability to support the reliability and reproducibility of experiments, through the collection and storage of provenance in an immutable way.

The SciPPGx workflow consists of 22 activities (tasks) in four sub-areas named, Area a (comparative genomics) with (SciHmm workflow [20]), Area B (phylogeny) with (SciPhy workflow [21]), Area C (evolutionary analysis) with (SciEvol workflow [22]), Area D structural bioinformatics analysis. Thus, for the workflow's execution, geographically distributed researchers belonging to different institutions participated in a collaborative experiment to discover new drug targets. To conduct the experiment, these researchers needed a collaborative environment that offers reliability in the provenance data, considering that reproducibility is essential in this context. These researchers also needed an environment that should offer interoperability in provenance data considering that the researchers execute parts of the experiment using different WfMS. The workflow was subdivided into three sub-workflows so that it could be executed collaboratively.

To collaborate, researchers need to specify a collaborative environment. For this purpose, the lead researcher of the experiment accessed the E-SECO

[9] https://nodejs.org/en/
[10] https://fabric-sdk-node.github.io/release-1.4/index.html
[11] https://www.mongodb.com/
[12] https://www.docker.com/
[13] https://yaml.org/
[14] https://docs.docker.com/compose/
[15] https://golang.org/

platform and specified the experiment's network. The blockchain network environment was specified in Amazon Elastic Compute Cloud (EC2) with three peer node (virtual machines). In each peer node was executed one workflow. In the virtual machine 1: SciHmm workflow [20], virtual machine 2: SciPhy workflow [21] and virtual machine 3: SciEvol workflow [22]. To collect the provenance data, each researcher, through a web service, collected the output and input data for each task in their sub-workflow. Then, the Wrapper layer maps the provenance data to the ProvONE model, as detailed in subsection 3.5. After being translated, these provenance data were sent to the Client layer, which sent it to the blockchain network as transactions through smart contracts (chaincode). The chaincode then stored the provenance data in the state DB and the blockchain file system, which was then shared between the blockchain network nodes. These steps provided integration, transparency (the data was visible to all researchers in the experiment), immutability, and trust for provenance data. To facilitate the analysis and understanding of the experiment's execution, the researchers could process queries to evaluate real-time provenance data. Figure 7 illustrates the user interface for executing queries against provenance data.

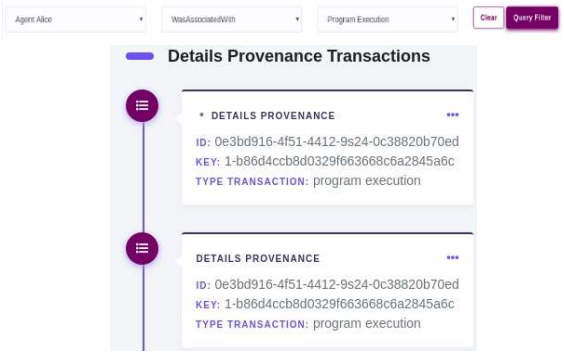

Figure. 7. Provenance query details.

4.2. Observed evidence and limitations

Some of the technological components identified as important could be processed and used, such as the provenance distributed management with blockchain processing. Initially, looking at the conduction of the feasibility study, the use of BlockFlow sounds promising. Besides, the implementation of the architecture is feasible. From the evidence presented, we confirmed the technical feasibility of the architecture. It offers components that can facilitate collaboration in scientific experimentation, considering scientific reproducibility and correct interpretation of scientific data among geographically distributed research. The technologies were sufficient for the operation of the architecture. One of the disadvantages and limitations when storing information in a blockchain-based application is that it is not possible to store image files, it is necessary to store hashes of information as detailed in subsection 3.5. Although this limitation can be overcome with IPFS, at BlockFlow, we still share all the in and out data generated during the execution of the collaborative workflow outside the chain. In this way, it is necessary to verify the data integrity, comparing if the stored hash corresponds with the data used as input and output during the execution of the workflow. In section 4.3, we had evidence that the system can operate with low latency, but the approach has not been compared with other approaches such as distributed databases. Therefore, during the feasibility study, some minor problems in the architecture could be identified and corrected. Besides, we provided some adjustments in the usability and organization of information. Thus, the results observed with this feasibility study enable and motivate the planning and execution of formal experiments using the architecture. In this vein, we are conducting an experiment considering patients' contamination by the new SARS-CoV-2 coronavirus.

4.3. Performance evaluation

In collaborative scientific workflow scenarios, researchers often store or simultaneously query the provenance repository, either to monitor it or to plan future actions. Thus, in this context, efficient mechanisms are required for both storage and query of provenance data. In this way, we evaluate our approach in terms of average throughput, varying the workload of transactions (10 to 10,000), between requests, (write/invoke and query) from provenance data, in the ledger performed by a set of peers (four peer) simultaneously, and batch size with 5tx. To evaluate we use an infrastructure of VM instances on Amazon Elastic Compute Cloud (EC2) running Ubuntu 16.04 and the Hyperledger Caliper benchmark. Table 1 shows the Average throughput per second for different numbers of transactions. After the analysis, we had evidence that the system can operate with a low latency. This result provides initial evidence that we can offer scalability and efficiency in distributed environments of scientific experimentation.

Table 1. Transactions Average Throughput

	Average Throughput per Second			
	10	***100***	***1000***	***10000***
Invoke	1.54 (s)	2.65 (s)	3.52 (s)	4.43 (s)
Query	0.01 (s)	0.01 (s)	0.08 (s)	1.39 (s)

5. Related work

Costa et al. [23] proposed an architecture that combines distributed workflow management techniques with provenance data management. Unlike blockchains, there is still a certain measure of centrality in traditional distributed architectures, leading to low reliability of provenance data. These systems also have security problems in information storage, considering that any authorized user can corrupt or alter the provenance data. Thus, it is necessary for the reproducibility and reliability of scientific experiments, that any user cannot alter the stored data. In this way, in BlockFlow, provenance data are stored immutably in the blockchain environment. Mendes et al. [6] proposed an architecture based on a Polystore approach to represent heterogeneous provenance data generated by different WfMSs, in a collaborative science scenario. Oliveira et al. [24] presented a proposal for integrating heterogeneous provenance data from distributed and heterogeneous workflows. However, these approaches have as the main disadvantage, a centralized storage system for provenance data. If the central server is compromised, the provenance data can be compromised and lost. Therefore, there is no single point of failure in the Blockchain architecture, once the provenance data is decentralized, shared among the geographically distributed researchers.

Several studies in the literature discuss the use of blockchain technology to enhance collaboration and reproducibility in e-science [25] [26]. In recent years, several works indicate this technology as promising for storing provenance data. Chen et al. [27] proposed a blockchain-based approach, named Prochain, to share provenance data from the execution of scientific workflows in a distributed community. However, the authors do not consider an environment where the provenance data is shared only between interested and duly authorized parties, like BlockFlow. Fernando et al. [28] proposed a blockchain-based system called SciBlock to provide tamper-proof and reputable storage for scientific workflow provenance data in a distributed collaborative environment. In SciBlock, the authors, in capturing and storing provenance, do not distinguish between prospective, retrospective, and evolutionary provenance. These provenance types have been identified as an essential requirement for every computational process in a workflow to achieve reproducibility. Thus, at BlockFlow, provenance data is integrated, stored, and shared among geographically distributed researchers through the ProvONE model, which considers the prospective, retrospective, and evolution provenance. Ramachandran et al. [29] proposed an architecture called Smartprovenance, based on blockchain for the safe and immutable management of provenance data. It tracks the changes in scientific documents on cloud platforms and records the provenance on the blockchain over updates to those documents, based on a voting mechanism. Smartprovenance uses the OPM provenance model and smart contracts to record provenance data in an immutable way. However, Smartprovenance does not use blockchain technology as distributed networks and does not has a real-time provenance data querying mechanism. Liang et al. [30] proposed a blockchain-based system called ProvChain to track provenance data in cloud storage applications. ProvChain tracks all changes in cloud storage applications and records each of these changes as provenance data in the blockchain. However, provenance data can be accessed by unauthorized users that belongs to the network.

6. Conclusions

This paper presented BlockFlow, a blockchain-based architecture where scientists can conduct experiments, share, and store provenance data through a trusted collaborative environment. The proposed solution is integrated into the Scientific Software Ecosystem Platform called E-SECO. We presented a feasibility study to enable researchers to collaborate for scientific experimentation in a trustful and transparent environment, sharing their data in an integrated manner. Thus, it could be noted that the solution leverages scientific collaboration by providing means of transparency, reliability, and reducing the heterogeneity of shared data in collaborative scientific workflows, as well as facilitating the interpretation and analysis of this data by geographically distributed researchers.

This work was developed to enhance the reproducibility, privacy, transparency, and interoperability in scientific software ecosystems platforms. Therefore, data provenance and blockchain provided and implemented through this approach are limited to this objective and cannot be generalized. However, the knowledge constructed and the results obtained can be transferred to other contexts. As future work, we intend to facilitate analysis and understanding of the experiment's execution, using dashboards and graphic illustrations. Additional evaluations of the architecture are also being conducted, involving projects related to control of epidemics (Sars-Cov-2) and integrating data from IoT devices related to oil prospecting.

Funding: This work was partially funded by UFJF/Brazil, CAPES/Brazil, CNPq/Brazil(grant: 311595/2019-7), and FAPEMIG/Brazil (grant: APQ-02685-17).

7. References

[1] TENOPIR, Carol et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PloS one, v. 10, n. 8, p. e0134826, 2015.

[2] Baker, M. (2016). 1500 scientists lift the lid on reproducibility. Nature News, 533(7604),452. DOI:10.1038/533452a

[3] Davidson, S. B., & Freire, J. (2008, June). Provenance and scientific workflows: challenges and opportunities. In Proceedings of the 2008 ACM SIGMOD international conference on management of data (pp. 1345-1350). ACM.G. DOI:10.1145/1376616.1376772.

[4] Moreau, Luc, and Groth, Paul. 2013. "Provenance: an introduction to PROV." Synthesis Lectures on the Semantic Web: Theory and Technology 3.4 pp 1-129.

[5] Coelho, Raiane, Braga, Regina, David, José, Dantas, Mário, Strole, Victor, Campos, Fernanda- "Blockchain for Reliability in Collaborative Scientific Workflows on Cloud Platforms" IEEE ISCC 2020.

[6] Mendes, Y., Braga, R., Ströele, V., & de Oliveira, D. (2019, May). Polyflow: A SOA for Analyzing Workflow Heterogeneous Provenance Data in Distributed Environments. In Proceedings of the XV Brazilian Symposium on Information Systems (p. 49). ACM. DOI 10.1145/3330204.3330259

[7] V. Freitas, J. M. N. David, R. Braga and F. Campos, "An Architecture for Scientific Software Ecosystem (In Portuguese)", Work. Distrib. Softw. Dev. Softw. Ecosyst. Syst., pp. 41, 2015.

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

[9] Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., & Paulson, P. (2008, June). The open provenance model: An overview. In International Provenance and Annotation Workshop (pp. 323-326). Springer, Berlin, Heidelberg. DOI 10.1007/978-3-540-89965-5_31

[10] Groth, P., and Moreau, L. (2013). PROV-Overview. An overview of the PROV Family of Documents. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

[11] Lim, Chunhyeok, et al. (2010) "Prospective and retrospective provenance collection in scientific workflow environments." 2010 IEEE International Conference on Services Computing. IEEE, 2010. DOI: 10.1109/SCC.2010.18

[12] Koop, David, and Juliana Freire. (2014) "Reorganizing workflow evolution provenance." 6th {USENIX} Workshop on the Theory and Practice of Provenance (TaPP 2014). 2014.

[13] Víctor Cuevas-Vicenttín, B Ludäscher, P Missier, K Belhajjame, F Chirigati, Y Wei, and B Leinfelder. (2015). ProvONE: A prov extension data model for scientific workflow provenance.

[14] 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. DOI /10.1007/978-3-319-39028-4_9

[15] 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. DOI 10.1145/3190508.3190538

[16] Lamport, Leslie. "Paxos made simple." ACM Sigact News 32.4 (2001): 18-25.

[17] Manikas, K. (2016). Revisiting software ecosystems research: A longitudinal literature study. Journal of Systems and Software, 117:84–103. DOI 10.1016/j.jss.2016.02.003

[18] Schaduangrat, N., Lampa, S., Simeon, S. et al. Towards reproducible computational drug discovery. J Cheminform 12, 9 (2020). DOI 10.1186/s13321-020-0408-x.

[19] J. Dias, D. de Oliveira, M. Mattoso, K. A. C. S. Ocana, and E. Ogasawara. Discovering drug targets for neglected diseases using a pharmacophylogenomic cloud workflow. In IEEE 8th Int. Conf. on E-Science (e-Science), pages 1–8, 2012.

[20] K.A.C.S. Ocaña, D. Oliveira, J. Dias, E. Ogasawara, and M. Mattoso, 2011, "Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific Workflow", In: 2011 IEEE e-Science, p. 190-197

[21] K.A.C.S. Ocaña, D. Oliveira, E. Ogasawara, A.M.R. Dávila, A.A.B. Lima, and M. Mattoso, 2011, "SciPhy: A Cloud-Based Workflow for Phylogenetic Analysis of Drug Targets in Protozoan Genomes", In: Advances in Bioinformatics and Computational Biology, p. 66-70.

[22] K.A.C.S. Ocaña, D. de Oliveira, F. Horta, J. Dias, E. Ogasawara, and M. Mattoso, 2012, "Exploring Molecular Evolution Reconstruction Using a Parallel Cloud-based Scientific Workflow", In: Advances in Bioinformatics and Computational Biology, p. 179-191.

[23] Costa, Flavio, Daniel de Oliveira, and Marta Mattoso. "Towards an adaptive and distributed architecture for managing workflow provenance data." 2014 IEEE 10th International Conference on e-Science. Vol. 2. IEEE, 2014.

[24] Wellington Oliveira, Paolo Missier, Kary Ocaña, Daniel de Oliveira, and Vanessa Braganholo. 2016. Analyzing provenance across heterogeneous provenance graphs. In IPAW. Springer, 57--70.

[25] Van Rossum, J. (2017). Blockchain for research. Science, November DOI 10.6084/m9. figshare

[26] Karastoyanova, D., & Stage, L. (2018, June). Towards collaborative and reproducible scientific experiments on blockchain. In International Conference on Advanced Information Systems Engineering (pp. 144-149). Springer, Cham. DOI 10.1007/978-3-319-92898-2_12

[27] Chen, W., Liang, X., Li, J., Qin, H., Mu, Y., & Wang, J. (2018, December). Blockchain Based Provenance Sharing of Scientific Workflows. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 3814-3820). IEEE.DOI:10.1109/BigData.2018.8622237

[28] Fernando, Dinuni, et al. "SciBlock: A Blockchain-Based Tamper-Proof Non-Repudiable Storage for Scientific Workflow Provenance." 2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC). IEEE, 2019.

[29] 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). CM DOI 10.1145/3176258.3176333

[30] 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 DOI: 10.1109/CCGRID.2017.8