SciLedger: A Scientific Workflow Provenance and Data Sharing Platform
IEEE Conference on Collaboration and Internet Computing ‘22

Hamilton Hardy  Reagan Hoopes  Min Long  Gaby Dagher

December 14, 2022
Table of Contents

1. Introduction
2. Related Work
3. Background
4. Architecture
5. Experimental Evaluation
6. Conclusion
INTRODUCTION
Motivation

- Scientific researchers collaborating from different locations
- Lack of way to ensure research integrity
  - 8.3% committed falsification/fabrication at least once from 2017-2020 [10]
- Increased requirements for data sharing from governmental and private funders [11]
- Flexibility within science
  - 60% of pre-established workflows concluded with null results [7]
  - Invalidation
Challenges

- Balancing contradictory needs of scientific research
  - Integrity limits flexibility
  - Public systems promote accessibility, but limit user privacy
  - Blockchain requires off-chain storage for scientific data which introduces security concerns
The Problem We Address

Scientific researcher’s needs for a system that:

- Is specific to scientific workflow provenance
- Allows for data sharing
- Supports complex processes such as branching and merging
- Provides a sufficient level of user privacy
Contributions

- The SciLedger system
- Public, blockchain-based platform that supports open-access data sharing and complex workflow operations
- Invalidation mechanism
- Implementation and experimental evaluation
RELATED WORK
Scientific Workflow Management Systems

- Kepler [2]
- Taverna [3]
- Galaxy [1]
- KNIME [4]
- Pegasus [5]

Key Features
- Locally Maintained Storage
- Scientific Field Specific
Generic Blockchain Solutions

- LineageChain [13]
  - Event Listeners for Data Modification
- BlockCloud [16][15]
  - Network Consensus by Staking cloud storage
- ProvHL [8]
  - Access Controls for Private Data
- Sifah et al. [14]
  - Data Ownership Permissions
- Key Features
  - Private Blockchains
  - Generic Solutions
Scientific Workflow Blockchain Solutions

- SmartProvenance [12]
  - Threshold Based Voting Smart Contracts
- Bloxberg [17]
  - Unique Provenance Model
- SciChain [6]
  - Optimized for High Performance Computing
- SciBlock [9]
  - Time Stamp Invalidation Mechanism
- Key Features
  - Private Blockchains
  - Limited in Features
BACKGROUND
Scientific Workflows and Provenance

- **Sci**
- **T**
- **d**

Graph showing the workflow with nodes and edges.
Merkle Trees

(a) Proving membership of data point
8

(b) Proving non-membership of data point 4
ARCHITECTURE
Overview

- Scientific Provenance Collection
- Complex Multi-Workflow System
- Dependency based Invalidation
- Two Tree Merkle Verification
### Scientific Provenance Collection (Cont.)

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task ID</td>
<td>The task’s assigned identifier value</td>
</tr>
<tr>
<td>Workflow ID</td>
<td>The workflow’s assigned identifier value</td>
</tr>
<tr>
<td>User ID</td>
<td>Public key belonging to the task performer</td>
</tr>
<tr>
<td>Submission Time</td>
<td>Submission time to the quorum</td>
</tr>
<tr>
<td>Input Data</td>
<td>Hash pointer to data before modification</td>
</tr>
<tr>
<td>Output Data</td>
<td>Hash pointer to data after modification</td>
</tr>
<tr>
<td>Valid Merkle Root</td>
<td>Top hash for valid Merkle tree</td>
</tr>
<tr>
<td>Invalid Merkle Root</td>
<td>Top hash for invalid Merkle tree</td>
</tr>
<tr>
<td>Other</td>
<td>Extra fields custom provenance values</td>
</tr>
</tbody>
</table>
Complex Multi-Workflow System

- Workflow Design
  - Merging
  - Branching
  - Multiple Workflows

- Inception Block
  - Predefined Workflow Design
  - Public Keys of Authorized Users

Figure: Sample SciLedger blockchain visualized as Workflows
Dependency Based Invalidation

- Invalidation Block
- Added to End of Workflow
- Updates Merkle Trees

Figure: Sample SciLedger blockchain visualized as Workflows
EXPERIMENTAL EVALUATION
Implementation

- Workflow Generator
  - Lorem Ipsum data
  - Branching and Merging Complexity
  - Valid and Invalid Merkle Trees
- Block Constructor
  - Provenance Record Construction
  - Transaction Header
- Blockchain
  - Node Consensus
Quorum Experiment Setup

- Malicious Activity in Scientific Research
  - 8.3% Maliciously Manipulated Data [10]
  - Fix Expected Malicious actors in the Network to be less than 12%.

- Parameters
  - Network Size (Scalability)
  - Quorum Size relative to the Network
  - Quorum Consensus Threshold
Quorum Parameter Experiment Results
Quorum Parameter Experiment Results

![Graph showing quorum parameter experiment results with bars for different quorum sizes and thresholds. The x-axis represents quorum size ranging from 0.05 to 0.25, and the y-axis represents good quorums ranging from 0% to 100%. The graph includes bars for quorum thresholds 0.7, 0.75, 0.8, 0.85, and 0.9.](image)
Quorum Parameter Experiment Results

![Graph showing the results of the quorum parameter experiment. The x-axis represents the quorum size (0.05 to 0.25) and the y-axis represents the percentage of good quorums. The graph includes bars for different quorum thresholds (0.7, 0.75, 0.8, 0.85, and 0.9).]
Additional Experiments in the Works

- Block Upload Speed
- Block Verification Transaction Analysis
  - Existence and Validity of Block
    - Valid Merkle Tree of Last Block Added
    - Valid Merkle Tree of the Block in the chain and absent from Invalid Merkle Tree of Last Block
  - Existence of Block
    - Valid Merkle Tree of the Block in the chain
    - Brute Force that recurses over chain until Block found
  - Non Existence of a Block
    - Absence from Valid Invalid Merkle Tree of Last Block
    - Brute Force that recurses over all blockchain until block is not found
CONCLUSION
Summary

- We propose SciLedger: a blockchain-based solution that supports open-access data sharing for scientific workflow provenance and complex workflow operations.
- We propose novel invalidation and merkle tree verification methods that allows researchers to modify workflows in a way that minimizes unnecessary repetition.
- SciLedger’s implementation shows such a system is possible.
- Experimentation proves our system’s scalability and efficiency.
Future Work

- Differential Data Privacy
- Consensus Mechanisms
- Scientific Data Verification in Blockchain
- Activity Privacy
The 8th IEEE International Conference on Collaboration and Internet Computing

December 14-16, 2022, Las Vegas, Nevada, USA (tentative)
Questions?


