A Blockchain based Data Production Traceability System

Research Project 2

Sandino Moeniralam
Wednesday February 28, 2018

University of Amsterdam
Introduction

- Need for data lineage
- Copernicus EO Sentinel-2 mission
- Blockchain based
Problem statement

- Reproducibility crisis
- Ideal situation
- Copernicus EO missions largest in history
- Version Control System insufficient
Reproducibility crisis

Figure 1: 1,500 scientists lift the lid on reproducibility
Source: https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
Related Work

Technologies

1. BigchainDB
2. Ethereum

Implementations

1. Provenance
2. Quality Assurance for Essential Climate Variables (QA4ECV)
3. VCS-Blockchain
What requirements should a Blockchain based production traceability system for satellite data adhere to?

- What does the data production process of Sentinel-2 Copernicus’s Earth Observation data look like?
- What types of data are to be distinguished?
- How does one capture all the steps of the data production process?
Data Lineage and Data Provenance

- Difference data lineage data provenance
- Several layers of abstraction
- Different views
- Open source provenance capture applications
Copernicus Sentinel-2 EO missions

- World’s largest single earth observation program
- Sentinel 1-7 planned
- 30 satellites in total
- Different companies involved including Airbus, EUMETSAT, SpaceX
Types of data

- The datasets themselves
- The production environment
  - Entire OS with applications
  - Python virtual environment
- The production process
  - Human view: comments, explanation
  - Machine view: automatic scripts
Satellite Data Processing Levels

- No strict definitions
- Level 0, 1A, 1B, 1C, 2A, 2B, 3A, 3B and 4
- Published from level 1C onwards
Blockchain

Advantages

• Immutable
• Distributed
• Secure
• Open

Disadvantages

• Scalability issues
• Computationally expensive

Figure 2: Distributed ledger
Source: https://elearningindustry.com/bitcoin-blockchain-impacting-elearning-industry
Data

- **Bitcoin**: transactions
- **Ethereum**: scripts
- **BigchainDB**: storage
Figure 4: Provenance Traceability Chain
Source: http://www.qa4ecv.eu
Proposed design

Blockchain data

- Cryptographic hash of the previous block
- Timestamp
- Proof-of-work

Data

- Hash(dataset)
- Pointer to dataset
- Hash(production environment)
- Pointer to production environment
- Hash(production process)
- Pointer to the production process
### Schematic sketch

#### Table 1: A schematic sketch

<table>
<thead>
<tr>
<th>Block 0</th>
<th>Block 1</th>
<th>Block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>hash(0)</td>
<td>hash(Block 0)</td>
<td>hash(Block 1)</td>
</tr>
<tr>
<td>timestamp</td>
<td>timestamp</td>
<td>timestamp</td>
</tr>
<tr>
<td>proof-of-work</td>
<td>proof-of-work</td>
<td>proof-of-work</td>
</tr>
<tr>
<td>hash(dataset V1)</td>
<td>hash(dataset V2)</td>
<td>hash(dataset V3)</td>
</tr>
<tr>
<td>pointer to V1</td>
<td>pointer to V2</td>
<td>pointer to V3</td>
</tr>
<tr>
<td>hash(PE #1)</td>
<td>hash(PE #2)</td>
<td>hash(PE #3)</td>
</tr>
<tr>
<td>pointer to PE #1</td>
<td>pointer to PE #2</td>
<td>pointer to PE #3</td>
</tr>
<tr>
<td>hash(PP #1)</td>
<td>hash(PP #2)</td>
<td>hash(PP #3)</td>
</tr>
<tr>
<td>pointer to PP #1</td>
<td>pointer to PP #2</td>
<td>pointer to PP #3</td>
</tr>
</tbody>
</table>
Discussion

- Volatile nature of digital data
- Production Environment large size
- Production Process complex
- Blockchain based
- Actual storage of the data unresolved
What requirements should a Blockchain based production traceability system for satellite data adhere to?

Every block should include the datasets, production environment and the production process for humans and machines.
Future Work

• More technical analysis into different Production Environments
• Ethereum Virtual Machine compatible
• Scalability issue
Questions?

“IT’S NOT BROKEN, IT’S A FEATURE...”

Sandino Moeniralam
sandino.moeniralam@os3.nl