A study on the Blockchain-powered Research Data Repository with FAIR Principles

Yeongheon Song*†, Minho Lee*†

* Research Data Sharing Center, Div. of National Science and Technology Data, Korea Institute of Science and Technology Information (KISTI)
† Dept. of Data and HPC Science, University of Science and Technology (UST)

Introduction and Problem Statement

- The need for *preserving research data* has been exponentially increased due to rapid increase in a computation capacity and open science policy.
- Open science itself is not a new concept, but *ensuring the reproductivity* of scientific data and *improving availability* is got more importance in the concept of open data.
- *Transparency of scientific research* also need to be ensured by registering all of research procedures because of the problem of research integrity.
  - Intentional or accidental corruption of research data
  - P-hacking: Removal of outliers / Clustering groups after experiments, Set hypothesis after result analysis

Blockchain

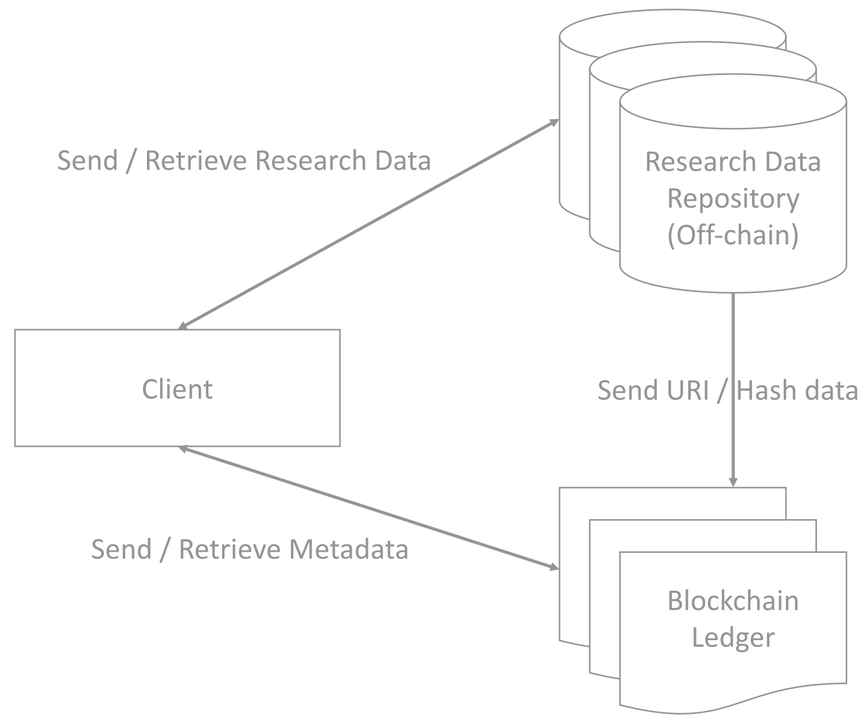
- As its name suggests, The data structure of Blockchain composed of a previous hash of the block and transaction data. Because hashes are *calculated sequentially*, if transaction data is compromised, the entire hash value after that block must be recalculated.
- Blockchain solutions are formed in type of public, private and consortium.
- Scientific blockchain solutions usually consist of *public* blockchain. (i.e., To promote the sharing of research data and exchange it with tokens)
- However, the use of public Blockchain and cryptocurrency could increase *uncertainty* and *degrade* the stability of the system.
  - The price of cryptocurrency is highly variable and the value of the token can be easily changed.
  - Consensus algorithms such as PoW(Proof of Work) requires a large number of calculation and profligate resources.
  - From the perspective of data privacy, smart contract in public network may leak researcher's personal information or unverified research data.

FAIR Principles

- As the volume of research data grows exponentially, data provider must ensure integrity and quality and promote the usability of research data.
- The FAIR guiding principles were established to remove the barriers to discovery and reuse and improve utilization of research data. FAIR principles consist of the following components.
  - *Findable*: (Meta)data are assigned with a UID (Unique Identifier) and registered in a searchable resource.
  - *Accessible*: (Meta)data are retrievable using a standardized communication protocol.
  - *Interoperable*: (Meta)data use formal vocabularies which meet FAIR Principles.
  - *Reusable*: (Meta)data are richly described with accurate and relevant attributes.

System Architecture

- We propose a conceptual model for research data repository based on Hyperledger Fabric, a *permissioned* Blockchain solution.



<Figure 1> Basic Architecture of Proposed Solution

- Each researcher registers his/her data in research data repository. But research data itself is not uploaded to Blockchain ledger, but *off-chain* storage (e.g., Content Delivery Network).
- This is because as the size of block which contains the data bigger, the time to propagate it to other nodes increases rapidly, which can cause bottlenecks in the entire Blockchain network.
- The storage calculates the hash value (e.g., MD5, SHA256) and merges it into transaction with the metadata and URI(Uniform Resource Identifier) and passes them into Blockchain ledger.
- The membership service provider(MSP) manages membership operation using the certificates used on a single channel. All peers in the network have its own identity and it is all known to MSP and coordinators. Unlike public blockchain, a transaction does not require additional computations in the verification process thanks to its consensus mechanism.

FAIR Principles and Blockchain Repository

- *Accessibility*: The pinpoint of Blockchain technology is decentralization. it can ensure that metadata is accessible in the repository, despite the data is corrupted. Even if a client is not connected to the web frontend server, it can still retrieve a metadata via RESTful API. An API involves user authentication procedures and it can be used to selectively provide information and to prevent unwanted information from being leaked.
- *Reusability*: Blockchain-based research can be used to track data manipulation and enable stakeholders to verify it. During the peer-review process when submitting a paper, the reviewer can check data repository and confirm analysis procedure. And if there is any data loss or corruption, researchers can figure out when it occurs. The system also can provide data uploaders with trust, since each download is recorded in a Block with a timestamp.

Conclusions

- We purposed a research data repository model based on a permissioned Blockchain platform, which helps scientists deal with research misconduct because it ensures transparency in transactions.
- However, There can be a single point of failure if the storage for research data is configured as only a single node. Therefore, Multi-node based models such as HDFS and Lustre or Blockchain-based storage models such as Ethereum Swarm are needed.
- Smart Contract and MSP in the Blockchain ledger also can be used to set data permissions for repository users. This allows data to be shared among researchers and it can be provided to reviewers for peer review purposes.
- This is a initial proof-of-concept model and we expect future studies would provide a detailed implementation of research data provenance and sustainability for long-term preservation through data quality control, peer management, and storage security systems.

Acknowledgements: This work formed part of research project carried out at the Korea Institute of Science and Technology Information (KISTI). (K-19-L01-C04-S01 Construction of Research Data Open Platform and its Utilization Support)

PRAGMA (Pacific Rim Applications and Grid Middleware Assembly) 36 Workshop: Jeju, South Korea. April 24-27, 2019

KiSTi www.kisti.re.kr Korea Institute of Science and Technology Information

UST 과학기술연합대학원대학교 UNIVERSITY OF SCIENCE & TECHNOLOGY