COMP4801

FINAL YEAR PROJECT

2017-2018

INTERMEDIATE REPORT

Data Marketplace powered by Blockchain

Topic
Blockchain Research and Implementation

Student

Lo Cheuk Yin 3035106243

Supervised by Dr. S.M. Yiu

Project Objective

The objective of this FYP project is to build a data marketplace powered with blockchain technology and decentralized storage. I envision the product of this project a catalyst for cooperative AI research and data sharing, and the hub for data exchange. Ultimately, it strives to increase and improve the quantity, quality and diversity of datasets available in the global lab of AI.

The objective is to encourage data owners to share their data. This is achieved by the adoption of blockchain technology, which ensures the data ownership of data, upholding the integrity of market participation history of data traders, and eliminate middleman risks. By making quality data available to AI researchers, they are powered to produce more powerful and better AI models, thus benefits the society.

Within the FYP period, I aim to finish an MVP (minimum viable product) of the data marketplace. The MVP will be composed of the followings:

- ✓ A web interface of the marketplace to view datasets, upload data and download data

- ✓ Smart contract logics in handling transactions on blockchain

- ✓ Decentralized storage system storing data

Project Background

Last year, I worked on an academic project that involves AI development. The client was trying to incorporate an AI into a stock price projection system. By the time when I was working on the project, I found the data gathering process very tedious and time consuming. There are insufficient data available, or data are scattered around which greatly hinder the development process. I realize the need for a global registry of datasets to facilitate development, and eventually gives me this idea to develop the Data Marketplace.

To generate good machine learning models, the quantity and quality of data are as important as the algorithm. Many machine learning planners find themselves insufficient of data, hindering the adoption of the technology in businesses. The lack-of-data obstacle is slowing down the research and development of machine learning, just like how the lack of computing power had for the past decades. In other words, today, one major key to build successful AI is the ability to gather data.

I believe the lack-of-data obstacle could be overcome through extensive cooperative research. However, traditional centralized server-client model is not suitable for data sharing and exchange because of the ambiguity in the ownership of data. Data sharers have to bear the middleman risks when using centralized solutions.

This motivates me to study the opportunity of blockchain technology in tackling the problem. Blockchain has the potential to eliminate the middleman risks and ensure complete data ownership by participants of the data exchange. It is believed that lowered risks and assured data ownership encourage data owners to share or sell their data. At the same time,

Therefore, this blockchain project, having a vision of a global community of researchers, businesses and individuals cooperating in the R&D and adoption of machine learning, is launched.

Project Methodology

As Blockchain is the core technology to be adopted in this project, the platform will be built using Ethereum, where smart contract will be designed and implemented to control transactions in data marketplace.

Since data exchange is the major activity carried out in this platform, there will be a huge amount of data transferring within the platform, and large chunk of data needed to be stored through this platform. To satisfy the principle of decentralization, these data will not be stored centralized in the platform, decentralized storage will be implemented.

The platform will not store data onto the blockchain, instead, large objects of data will be stored through distributed storage to minimize cost to put data on blockchain. Currently, there are protocols and services available in the market for distributed object storage and distributed file system.

The considered system architecture is comprised of three layers: client layer, consensus layer and data storage layer.

Client Layer

The major responsibility of this layer is to govern the interactions between users and the system, which mainly serves as a user interface. There will be a web interface for users to interact with the system.

Consensus Layer

Consensus Layer is mainly implemented on Ethereum, which is on the blockchain through smart contracts

A smart contract is dedicated to Marketplace. The major functions are as follows:

- ✓ acting as a registry of data for sale;

- ✓ acting as a "middlemen" to settle transaction;

- ✓ registry of rating and comments.

Data Storage Layer

This layer is responsible for data storage, where data are transferred to this layer through consensus layer and the client layer. The technologies will be used in this project in this layer are BigchainDB and IPFS.

BigchainDB

It serves mainly as a database that shares properties of blockchain with transparency and immutability. The main reason to adopt this technology instead of directly writing data onto blockchain is to minimize cost.

IPFS

This is a protocol to serves as blob storage to store files and data in a distributed manner with nodes on the network.

What has been accomplished

- ✓ Preliminary Studies on:

    - Implementation method on Ethereum Smart Contract

    - Technology feasibility to be adopted in this project

- ✓ API connecting distributed storage using IPFS

What will be done

- ✓ Web interface for users to interact with the system

- ✓ Ethereum Smart Contract to handle transactions and system logics

- ✓ System Testing and Evaluation