Visvesvaraya Technological University, Belagavi

A Project Synopsis On

“Blockchain based data pipeline to ensure fairness in datasets crowd sourcing”

Submitted By

Vignesh T D
(4AD22CD056)

Ranjitha
(4AD22CD040)

Ananya Raj
(4AD22CD005)

Bharath N
(4AD22CD007)

Under the Guidance of

Mrs. Madhu Nagraj

Assistant Professor,
Department of CSE- Data Science

ATME College Of Engineering
13th KM Stone, Mysuru – Kanakapura – Bengaluru Road,
Mysuru - 570028

PipeChime: Blockchain based data pipeline to ensure fairness in datasets crowd sourcing.

Abstract

As artificial intelligence continues to shape modern life accelerated by its mainstream boom in recent years the demand for high-quality, diverse, and transparent datasets has never been greater. Yet, the infrastructure needed to gather and share such data remains largely controlled by centralized entities, creating barriers for broader innovation and collaboration.

This project proposes **PipeChime**, a decentralized framework that reimagines pipeline for AI model training. By leveraging blockchain's transparency and trust less architecture, PipeChime enables a more equitable and collaborative way to crowdsource, verify, and access datasets. The system integrates on-chain mechanisms for immutable metadata tracking, smart contract-based quality control, and a tokenized reward system that incentivizes high-quality contributions. Meanwhile, scalable offchain storage ensures the framework remains efficient and accessible.

I. Introduction

Artificial intelligence has been becoming a part of the daily life of human beings since the AI trend boom in 2020 [1]. It is becoming easier for humans and machines to communicate, enabling AI users to accomplish more with greater proficiency. AI is projected to add USD 4.4 trillion [2] to the global economy through continued exploration and optimization [3]. This continuous learning and exploration with the machine requires immense volume of quality data to streamline the demanding requirements and processes.

For instance, state-of-the-art image classification models like ResNet-50 [4] are designed to be trained on high-end GPUs on the ImageNet dataset [5], which has over 14 million hand-annotated images. Such resource requirements have led to the concentration of AI development, which creates an obstacle to innovation by the masses.

There exists Decentralized AI marketplaces (such as SingularityNET [6] and Ocean Protocol [7]) that have emerged as promising solutions to these challenges. These platforms enable people and organizations to pool their data and computing power for building and training machine learning models [12]. Utilizing blockchain technology, they support a secure, transparent, and reward-driven environment for accessing models and datasets in a decentralized manner, fostering both innovation and cooperative development.

In this synopsis, we will be proposing a blockchain based framework for data pipeline to seamlessly crowd source data required for the AI ecosystem called as “PipeChime” using an optimal blockchain technology required for our necessities among Algorand, Ethereum, Polygon – PoS (Eth L2), Solana etc. Users can ingests consensus data into the system through the ingestion layer and an on-chain protocol tracks the meta data and labels the input, The processed data is posted to an off-chain storage, facilitated with the interference of smart contracts layered with an auditing on top of it. A access layer is the endpoint where data consumers can get data once validated. To make the process fair and motivating, contributors could be rewarded based on quality of the data with a portion of the value as tokens provided by the data consumer in exchange for their input.

II. Purpose

The goal of this project is to build PipeChime, a practical framework that helps people contribute and access high-quality datasets for training AI models. Instead of relying on a few major players to provide all the data and compute power, this system opens the door for anyone to take part securely and fairly by using blockchain as the foundation.

PipeChime is designed to make the entire data pipeline more transparent and decentralized. With support from technologies like Algorand, Polygon (Ethereum Layer 2), and Solana, the platform allows users to upload data, which gets tagged and logged in a tamper-proof way. A smart contract layer keeps track of contributions, verifies data quality, and manages rewards. Validated data is then stored off-chain and made available to consumers through an access layer.

In notable terms, PipeChime ensures:

- Transparent and immutable metadata tracking
- Smart contract-based auditing and quality control
- Token-based reward distribution to contributors based on data quality
- Secure and scalable off-chain data storage access for consumers

In the process, Contributors are recognized and compensated in tokens, based on how valuable or accurate their data is. Overall, PipeChime offers a cleaner, fairer system for building machine learning models one that encourages open collaboration, protects data integrity, and lowers the barrier to entry for developers and researchers everywhere.

III. Objectives

- **Enable Blockchain Based Data Contribution**
  Create a platform where individuals and organizations can contribute AI

training data from anywhere, without relying on centralized gatekeepers. •

Ensure Data Traceability and Transparency

Use blockchain technology to immutably log metadata about each data contribution including origin, timestamp, and quality metrics to foster accountability and trust.

- **Design a Fair and Incentive-Based Reward Mechanism**
  Implement a token-based system that fairly compensates contributors based on the quality and relevance of their data, encouraging long-term participation and data integrity.

IV. Literature Review

Wikipedia: History of Artificial Intelligence [1]

This article provides a comprehensive timeline of artificial intelligence, from its conceptual origins to its modern breakthroughs. The section covering developments post-2022 highlights the surge in AI adoption, driven by models capable of performing tasks that were previously exclusive to human cognition. It marks the release of widely adopted generative tools as a key turning point in AI's integration into daily life and practical applications.

Forbes – "AI Could Add $4.4 Trillion to Global Economy" [2]

This piece presents a strong economic perspective on the impact of AI, forecasting its transformative role across industries. It emphasizes the financial value AI is expected to bring, largely due to its ability to optimize processes, reduce costs, and drive innovation at scale. The report also hints at structural challenges, particularly the access to high-quality data and the resources required to develop sophisticated AI models.

IBM – "Top AI Trends to Watch" [3]

IBM outlines emerging trends in artificial intelligence that are shaping its future. The article discusses the shift towards responsible and transparent AI, the growing need for explainability, and the evolution of edge computing. These trends suggest a focus on trust, data integrity, and decentralized processing, pointing toward a future where AI systems must be not only powerful but also accountable and fair.

Tao et al. – "Image Super-Resolution Using Deep Recursive Residual Networks" [4]

This technical paper explores a specific use case of AI in computer vision. It demonstrates how deep learning models, when trained with sufficiently large and high-quality datasets, can achieve state-of-theart performance in image enhancement tasks. The research underscores the importance of both computational power and dataset richness in training robust AI systems.

ImageNet – Large Scale Visual Recognition Dataset [5]

ImageNet has been a cornerstone in the field of machine learning, particularly for training and benchmarking computer vision models. With millions of annotated images, it illustrates how large, structured datasets have historically played a key role in the development of advanced AI models. However, its creation and maintenance also highlight the intensive human and computational resources required for such datasets.

Montes & Goertzel – "Decentralized AI: Bringing the Community In" [6]

This paper presents a theoretical foundation for decentralized artificial intelligence, advocating for systems that reduce central control over data and computation. The authors argue that democratizing AI can lead to more equitable innovation and increase public trust in intelligent systems. They also discuss how such models could evolve collaboratively across distributed networks.

Ocean Protocol – Data Exchange for Web3 [7]

Ocean Protocol provides a practical example of how blockchain can be used to facilitate secure and traceable data exchange. The protocol enables individuals and institutions to monetize and control their data while contributing to larger ecosystems. It promotes transparency and fairness, especially in datacentric applications like machine learning and AI model training.

Algorand [8]

Algorand is a permissionless, pure proof-of-stake blockchain platform designed for high throughput and minimal transaction latency. Its architecture enables fast finality, making it well-suited for applications that demand speed and low operational friction. Algorand prioritizes decentralization and aims to support scalable solutions without compromising security, while also maintaining relatively low energy consumption compared to traditional blockchains.

Ethereum (Layer 1) [9]

Ethereum is a decentralized, open-source blockchain known for its smart contract functionality and extensive developer ecosystem. As the first platform to introduce programmable contracts, Ethereum has become the foundation for many decentralized applications (dApps). However, due to its popularity and base-layer design, it often experiences scalability challenges, such as network congestion and high gas fees—issues that have driven interest in scaling solutions like Layer 2s.

Polygon – PoS (Ethereum Layer 2) [10]

Polygon (PoS) functions as a scalable Layer 2 chain running parallel to Ethereum. It uses a proof-of-stake consensus mechanism to offer higher throughput and significantly lower transaction fees. Polygon effectively bridges the Ethereum ecosystem with more scalable, low-cost solutions, enabling developers to leverage Ethereum’s security and compatibility while mitigating congestion and cost concerns. It is frequently adopted for decentralized finance (DeFi), gaming, and NFT-related platforms.

Solana [11]

Solana is a high-performance blockchain that uses a unique consensus mechanism combining proof of stake and proof of history. This architecture allows it to process thousands of transactions per second with minimal fees. Solana focuses on speed and efficiency, making it attractive for high-frequency applications like decentralized exchanges, real-time games, and interactive dApps. However, its rapid growth has also raised concerns around decentralization and network stability during peak usage.

V. Methodology

The following figure 1.1 illustrates a modular system designed to streamline and secure the process of crowdsourcing datasets for AI training. It involves Data Ingestion Layer, Meta Data tracking and quality evaluation, Off-chain storage, Smart contracts, Auditing and Access points.

1. Data Ingestion Layer

This is where contributors interact with the system. At this point, users submit datasets whether images, text, or structured data through an interface that collects and prepares the data for further processing. The layer performs basic validation such as format checks, duplicate detection, and structural compliance before passing the data onward. The goal here is to act as the first filter to minimize noise and maintain consistency.

2. On-Chain Metadata Tracker & Quality Evaluation

Once the data clears the ingestion phase, it is routed through a logic layer that assigns metadata, such as contributor identity (hashed), timestamps, and descriptive tags. This metadata is stored immutably on the blockchain to ensure traceability and tamper-resistance.

In parallel, the system performs a lightweight quality evaluation this might involve peer ratings, automated content checks, or voting mechanisms depending on the data type. The metadata and evaluation results together form a ledger entry that can be queried, audited, or referenced later.

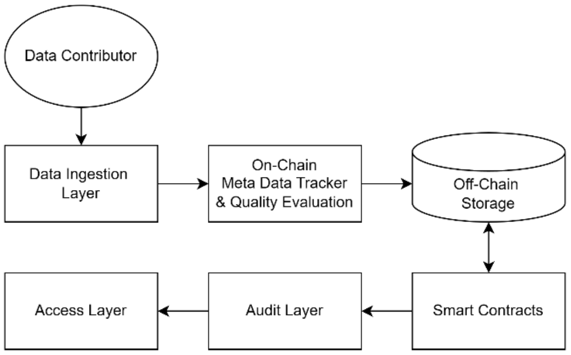

Fig 1.1 : Representation of Blockchain based data pipeline for fair datasets crowd sourcing

3. Off-Chain Storage

Due to the size and nature of dataset files, raw data is not stored directly on the blockchain. Instead, it is sent to a secure, distributed storage layer. Here, each data entry is hashed, and only the reference (or pointer) is kept on-chain. This approach keeps the system efficient while maintaining verifiability the hash ensures that any tampering with the stored data can be detected.

4. Smart Contracts

Smart contracts govern the automation within the system. These autonomous scripts are responsible for executing predefined rules without human intervention. They handle reward distribution, enforce quality thresholds, manage permissions, and respond to events like successful data verification or requests from consumers. Their deterministic nature ensures fairness and consistency across the board.

5. Audit Layer

Serving as the trust anchor, the audit layer continuously monitors system activities. It checks if data contributions match their on-chain records, verifies if smart contracts are performing as expected, and ensures no bypass or manipulation has occurred. This layer can generate reports or alerts for anomalies, providing a secondary shield against fraud or errors.

6. Access Layer

This final module enables authorized consumers AI researchers, developers, or organizations to query, filter, and access the validated datasets. Since all interactions are logged and tied to metadata, users can make informed decisions about data quality and provenance. The access layer ensures that only data with verified quality and approved metadata can be retrieved, closing the loop between crowd-contributed data and real-world AI applications.

This architecture, when combined with community participation and clear incentives, builds a reliable and fair ecosystem for sourcing AI training data one where quality, trust, and openness are built in from the start.

VI. Implementation

1. Core CLI / SDK

Languages: Python (using argparse, click, or typer), Go, Rust, or Node.js. **Purpose:** Acts as the main user interface, allowing users to run commands like:

```
myframework configure --api-key <key>
myframework data submit --file data.csv
myframework access request --dataset-id 123
```

Configuration: Stores API keys, blockchain URLs, and wallet info in config files or environment variables.

2. Data Ingestion (via CLI)

Submit Command: Packages data and metadata and sends them to the backend.

Backend API: REST or application APIs (built with Node.js, Python, Go, etc.) handle data upload and metadata storage.

Processing: Uses message queues and triggers smart contracts and quality checks.

3. Blockchain Interaction & Metadata Tracking

4. Off-Chain Storage

Storage Systems: IPFS, AWS S3, Google Cloud, Azure.

Upload/Download: Backend manages uploads; CLI can handle downloads via secure links or tokens. **5. Smart Contracts**

Function: Manage rules, access control, and logs on-chain.

Accessed Via Backend: Triggered by user commands via API interactions.

6. Audit and Access Layer

Audit Logs: CLI can fetch blockchain-based logs.

Example:

```
myframework audit log --dataset-id 123
```

Access Management: CLI commands manage access requests and data downloads.

VI. Expected Outcomes

- A working decentralized platform where anyone can contribute AI training data securely and transparently.
- Clear, traceable records for every piece of data added   making the entire process more accountable.
- A fair reward system that encourages high-quality contributions through token-based incentives.
- Reliable off-chain storage connected to an on-chain metadata system for better scalability.
- Smart contracts that handle data validation, reward distribution, and access permissions automatically.

VII. References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. IEEE, 2009

[2] Vitalik Buterin , Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform. 2014

[3] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3147–3155, 2017

[4] Gabriel Axel Montes and Ben Goertzel. Distributed, decentralized, and democratized artificial intelligence. Technological Forecasting and Social Change, pages 141:354–358, 2019.

[5] Trent McConaghy. Ocean protocol: Tools for the web3 data economy. In Handbook on Blockchain, pages 505–539. Springer, 2022

[6] IBM E-Article, The future of AI: trends shaping the next 10 years

[7] Forbes E-Article, Harnessing Generative AI: A $4.4 Trillion Opportunity For The Global Economy

[8] Empowering collaboration and data accessibility for AI in a decentralized blockchain based marketplace

[9] Anatoly Yakovenko, Solana: A new architecture for a high performance blockchain v0.8.13

[10] Mihailo Bjelic Sandeep Nailwal Amit Chaudhary Wenxuan Deng, POL: One token for all Polygon chains,

[11] Jing Chen, Stony Brook University, Algorand*

[12] Wikipedia E-Article, History of Artificial Intelligence, AI Boom