VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi-590018

A Project Report

on

"Pipechime: Analytics Tool for Federated Learning Based Machine Learning and Data Analytics"

BACHELOR OF ENGINEERING DEGREE
In
COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

Submitted by

Ananya Raj M.N	**4AD22CD005**
Bharath N	**4AD22CD007**
Ranjitha	**4AD22CD040**
Vignesh T.D	**4AD22CD056**

Under the guidance of

Mrs. Madhu Nagraj
Assistant Professor
Department of CSE (Data Science)

ATME College of Engineering,

13th Kilometer, Mysore-Kanakapura-Bangalore Road
Mysore-570028

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Belagavi-590018

ATME College of Engineering

13th Kilometer, Mysore-Kanakapura-Bangalore Road

Department of CSE (DATA SCIENCE)

CERTIFICATE

This is to certify that the project work entitled "Pipechime: Analytics Tool for Federated Learning Based Machine Learning and Data Analytics" is the bonfide work carried out by following students,

Ananya Raj M. N	**4AD22CD005**
Bharath N	**4AD22CD007**
Ranjitha	**4AD22CD040**
Vignesh T. D	**4AD22CD056**

In partial fulfillment for the award of degree of Bachelor of Engineering in CSE (Data Science) from the Visvesvaraya Technological University, Belagavi during the year 2024-25. It is certified that all the corrections or suggestions indicated for internal assessment have been incorporated in the Major-project report deposited in the department library. The Major-project report has been approved and satisfies the academic requirement with respect to Major-project work prescribed for Bachelor of Engineering degree.

Signature of the guide	Signature of HOD
Mrs. Madhu Nagraj	Dr. Anitha D B
Assistant Professor	Associate Prof. & HOD

ACKNOWLEDGEMENT

The successful completion of our Major project phase-1 would be incomplete without mentioning the names of the people who have made it possible. We are indebted to several individuals who have helped us to complete the project report.

We are thankful to **Dr. L Basavaraj**, **Principal**, ATME College of Engineering, for having granted us permission and extended full use of the college facilities to carry out this project successfully.

We express our profound gratitude to **Dr. Anitha DB & HOD & Associate Professor**, Department of CSE (Data Science) for her consistent co-operation and support.

At the outset we express our profound gratitude to our guide **Mrs. Madhu Nagraj, Assistant Professor**, Department of CSE (Data Science) for her consistent co-operation and support.

We are greatly indebted to our project coordinator Dr. Vinod Kumar. P**,** Associate Professor, Department of CSE (Data Science) for his timely inquiries into the progress of the project.

Lastly, we would like to thank our family and friends for their cooperation and support for success full completion of our project.

Ananya Raj M. N	**4AD22CD005**
Bharath N	**4AD22CD007**
Ranjitha	**4AD22CD040**
Vignesh T. D	**4AD22CD056**

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 federated framework that reimagines the data pipeline for AI model training. By leveraging federated architecture and compliance, PipeChime enables a more equitable and collaborative way to crowdsource, verify, and access datasets.

As artificial intelligence (AI) becomes increasingly embedded in the fabric of modern society, the role of data in shaping AI outcomes has grown more critical than ever. The recent mainstream boom in AI technologies from generative models to decision-making systems has further amplified the demand for datasets that are not only large and diverse but also high in quality, transparent in origin, and ethically sourced.

However, the current data infrastructure powering AI development remains predominantly centralized. Large corporations and data monopolies control the majority of data collection, curation, and distribution processes.

CONTENTS

	Page No.
CHAPTER 1: INTRODUCTION	
1.1 Preamble	1-2
1.2 Motivation	3
1.3 Problem Statement	3
1.4 Objectives of the project	4
CHAPTER 2: LITERATURE SURVEY	
2.1 Introduction	5-6
2.2 Related works/Literature review	7-10
2.3 Outcome of the literature review	11
CHAPTER 3: APPLICATION AND REQUIREMENTS	
3.1 Applications	12
3.2 Requirements	13
Conclusion	14
Reference	15-16

CHAPTER 1

INTRODUCTION

1.1 Preamble

Artificial intelligence has been becoming a part of the daily life of human beings since the AI trend boom in 2020. 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 to the global economy through continued exploration and optimization. This continuous learning and exploration with the machine require immense volume of quality data to streamline the demanding requirements and processes.

For instance, state-of-the-art image classification models like ResNet-50 are designed to be trained on high-end GPUs on the ImageNet dataset, 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 Singularity NET and Ocean Protocol) 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. 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 , 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. An 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.

This project aims to design a blockchain-based data pipeline that upholds fairness in crowdsourced dataset collection, paving the way for more trustworthy and inclusive artificial intelligence.

To address these challenges, blockchain technology offers a decentralized, transparent, and tamper-proof solution. By integrating blockchain into the data pipeline, we can ensure that all data contributions are traceable, fairly rewarded, and securely managed.

In today's digital world, many artificial intelligence systems rely on large amounts of data to work effectively. Crowdsourcing is a popular way to collect such data from many people across different regions. However, this method often raises concerns about fairness, transparency, and bias.

Blockchain technology offers a way to solve these problems. It allows data to be stored in a secure, transparent, and unchangeable way. By using blockchain in the data collection process, we can make sure that every data contribution is recorded, verified, and rewarded fairly.

1.2 Motivation

As highlighted in the reviewed literature, artificial intelligence has rapidly evolved from a niche field to a transformative force across industries. With this growth, the demand for high-quality, transparent**,** and diverse datasets has surged, as noted by the reliance of deep learning models on large-scale annotated datasets like ImageNet. However, existing AI data pipelines are primarily controlled by centralized institutions, which limits broader participation, reduces transparency, and introduces inherent biases in data availability and access.

Objective:

The main goal of PipeChime is to create a fair and decentralized system for collecting and sharing AI training data. It allows anyone to contribute datasets without relying on big companies. Every data entry is tracked on the blockchain to ensure transparency and trust. Smart contracts are used to check the quality of the data and to reward contributors based on how useful their data is. To handle large files efficiently, the actual data is stored off-chain while the important details are kept on-chain.

1.3 Problem Statement

Artificial intelligence needs large and good-quality datasets to work properly. But today, most of this data is controlled by big companies. This makes it hard for small developers and researchers to get the data they need Also, the sources of data are often not clear or trustworthy. People don’t know where the data comes from or if it’s reliable. There is no proper system to check the quality of the data Another issue is that people who share useful data don’t get any rewards or recognition. This discourages others from contributing and slows down progress in AI. To solve these problems, we need a new system that is open, fair, and rewards contributors. PipeChime aims to build that system using blockchain technology to make data sharing transparent, secure, and equal for everyone.

1.4 Objective of the Project

- **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

The primary objective of this project is to design and implement a decentralized**,** transparent, and fair data crowdsourcing pipeline using block chain technology. The aim is to ensure ethical data collection**,** fair participation and bias mitigation in datasets that are used to train AI and machine learning models.

Key Objectives:

1. **Ensure Data Fairness and Diversity**
   - Prevent dataset bias by enforcing demographic balance and representation across data contributions
2. **Establish Transparent Data Provenance**
   - Use block chain to record immutable logs of data origin, submission metadata, contributor identity (pseudonymous), and consent.
3. **Create a Decentralized Incentive Model**
   - Reward contributors and validators with tokens based on the quality and diversity value of the data they provide or verify.

CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

Artificial Intelligence has grown rapidly in recent years and is now used in many areas like healthcare, finance, education, and more. This growth depends heavily on large amounts of high-quality data to train accurate and reliable AI models. However, collecting and sharing such data is often managed by large companies, which creates barriers for students, researchers, and small developers who want to contribute to or benefit from AI development. Centralized data control also raises concerns about bias, lack of transparency, and limited access.

The literature reviewed for this project highlights both the challenges and opportunities in this area. Research shows that decentralized systems, such as those using blockchain technology, can help solve these issues by making data sharing more open, fair, and trustworthy. Projects like Ocean Protocol and SingularityNET have shown how blockchain can be used to create secure and reward-based data sharing platforms. This background supports the need for a new approach, which our project, PipeChime, aims to provide.

PipeChime will use blockchain to create a transparent and fair system where anyone can contribute data, ensure its quality, and get rewarded. This literature survey helps us understand the current problems and existing solutions, and guides us in building a better data pipeline for AI.

Many studies and articles point out that the growth of AI is heavily dependent on large and high-quality datasets. For example, models like ResNet-50 used for image recognition need millions of well-labelled images like those in ImageNet to perform well. However, collecting such huge datasets requires a lot of time, money, and human effort, which only big organizations can usually afford. This creates a gap where smaller developers and researchers are left behind.

Literature survey:

Artificial Intelligence has become a powerful tool used across various fields like healthcare, education, finance, and transportation. According to the article from *Forbes*, AI is expected to add over $4.4 trillion to the global economy due to its ability to improve productivity and innovation. However, this progress depends greatly on access to high-quality and diverse datasets. The ImageNet project, for example, showed how a large dataset with millions of labelled images helped improve AI in computer vision. But creating such datasets requires heavy resources and is often controlled by large tech companies, making it difficult for smaller developers and researchers to contribute or benefit.

Several researchers and platforms have proposed decentralized solutions to overcome these challenges. Montes and Goertzel highlighted in their work the importance of a community-driven AI system that removes centralized control. Similarly, *Ocean Protocol* and *SingularityNET* introduced blockchain-based platforms where people can share and access datasets securely and fairly. These platforms make use of blockchain's key features like transparency, immutability, and decentralization, and use smart contracts to ensure fair transactions. These systems show that it is possible to create open environments where data contributors are rewarded and trusted data is made available for all.

2.2 Literature review

Foundational Concepts

AI models rely heavily on the volume and quality of training data. The seminal ImageNet project by *Deng et al.* [5] demonstrated how large-scale annotated datasets significantly boost model performance, especially in computer vision.

Similarly, *Tai et al.* [4] showed that deep learning models for image enhancement like super-resolution networks perform best when trained on rich and high-resolution data. These studies affirm that high-quality datasets are the backbone of effective AI systems.

However, curating such datasets traditionally involves centralized, resource-intensive efforts, which raises the need for collaborative and distributed data pipeline

Recent Analyses and Critiques

Recent studies highlight both the potential and challenges of using blockchain in AI data pipelines. While platforms like Ocean Protocol and Singularity NET showcase real-world adoption of decentralized data exchange, critiques point to issues such as scalability, energy consumption (in Layer 1 chains), and difficulty in enforcing data quality. Furthermore, many token incentive systems focus on participation rather than quality, which may lead to data redundancy or manipulation. The need for better auditability, decentralized governance, and fairness in contributor recognition remains a recurring theme in contemporary literature.

Scalability concerns with Ethereum Layer 1 due to high gas fees and latency.

- Data quality enforcement is still immature in many decentralized platforms.
- Token rewards often incentivize quantity over data relevance and utility.
- Smart contract complexity can hinder adoption by non-technical users.

Audit layers are often centralized or semi-automated, reducing true decentralization

Applications

- PipeChime can serve as a trusted backend for AI developers to access diverse, validated datasets for tasks like image recognition, NLP, and predictive analytics.
- Crowdsourcing anonymized patient data with blockchain-backed traceability can support disease prediction, drug discovery, and public health analytics.
- Sensor data from various contributors (traffic, pollution, energy usage) can be securely logged and monetized via PipeChime for urban planning and real-time decision-making.
- Platforms like SingularityNET or Ocean Protocol can integrate PipeChime to enhance dataset acquisition with embedded quality checks and tokenized incentives.
- Farmers can upload crop data, soil quality metrics, and climate conditions, which can be accessed by AI systems for yield prediction and supply chain optimization.
- Universities and research labs can share experimental datasets, ensuring integrity and fair credit via on-chain contributor tracking and rewards.
- Crowdsourced economic indicators, transaction data, and sentiment analysis can be stored securely for use in algorithmic trading and fraud detection models.
- Labelled image, video, and text datasets sourced from volunteers can be validated.

Wikipedia E-Article, History of Artificial Intelligence, AI Boom The **history of Artificial Intelligence (AI)** has witnessed several significant periods of optimism and innovation, often referred to as "AI booms." These booms are characterized by rapid advancements, increased funding, and heightened public and academic interest [1].

Forbes E-Article, Harnessing Generative AI: A $4.4 Trillion Opportunity For The Global Economy. According to a Forbes article and a McKinsey Global Institute report, Generative **AI** (GenAI) has the potential to contribute up to $4.4 trillion annually to the global economy. This explosive growth is being driven by GenAI's ability to enhance productivity, transform industries, and create new value in ways traditional AI could not [2].

IBM E-Article, The future of AI: trends shaping the next 10 years Artificial Intelligence (AI) is entering a new era of maturity and integration, with IBM projecting that the next decade will be defined by human-centered AI**,** enhanced governance**,** and industry-specific innovations**.** AI will increasingly move from experimentation to deep, strategic deployment [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. The paper introduces the Deep Recursive Residual Network (DRRN) for single image super-resolution (SISR)—a task where the goal is to reconstruct a high-resolution (HR) image from a low-resolution (LR) input. The DRRN achieves state-of-the-art performance by combining recursive learning with residual learning, enabling deep network design without significant increases in parameters [4].

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 The paper introduces ImageNet**,** a large-scale, richly annotated, hierarchical image database that has become foundational to modern computer vision research. Designed to facilitate the development of more accurate and scalable visual recognition systems, ImageNet contains millions of labeled images organized according to the WordNet hierarchy [5].

Gabriel Axel Montes and Ben Goertzel. Distributed, decentralized, and democratized artificial intelligence. Technological Forecasting and Social Change, pages 141:354–358, 2019. This paper explores the emerging paradigm of Distributed, Decentralized, and Democratized Artificial Intelligence (D3AI)—a vision for AI development that is open, collaborative, and not controlled by centralized corporate or governmental entities. The authors argue that the future of AI must be inclusive, transparent, and equitably accessible [6].

Trent McConaghy. Ocean protocol: Tools for the web3 data economy. In Handbook on Blockchain, pages 505–539. Springer, 2022 In this chapter, Trent McConaghy presents Ocean Protocol**,** a decentralized data exchange platform designed for the Web3 economy**.** Ocean Protocol enables secure, privacy-preserving, and traceable data sharing by leveraging blockchain technology. It aims to unlock the value of data by connecting data providers and consumers while maintaining control, transparency, and monetization capabilities [7].

Jing Chen, Stony Brook University, Algor and Algor and is a next-generation blockchain platform designed to overcome the trilemma of scalability**,** security, and decentralization. Jing Chen's contributions focus on the theoretical underpinnings of Algorand's Pure Proof-of-Stake (PPoS) consensus mechanism, providing a mathematically robust foundation for building a fast, secure, and fair decentralized system [8].

Vitalik Buterin, Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform 2014. In this seminal white paper, Vitalik Buterin proposes Ethereum, a decentralized platform that expands the use of blockchain technology from simple cryptocurrency (like Bitcoin) to support smart contracts and decentralized applications (dApps). Ethereum aims to serve as a global, trust less computing platform, offering a flexible and programmable blockchain [9].

Mihailo Bjelic Sandeep Nailwal Amit Chaudhary Wenxuan Deng, POL: One token for all Polygon chains This paper introduces POL**,** a next-generation token designed to serve as the single native asset for all Polygon chains**,** including PoS, zkEVM, and future Layer-2 chains [10].

Anatoly Yakovenko, Solana: A new architecture for a high performance blockchain v0.8.13 In this paper, Anatoly Yakovenko introduces Solana**,** a high-performance Layer 1 blockchain designed for scalability without sacrificing decentralization or security [11].

Empowering collaboration and data accessibility for AI in a decentralized blockchain based marketplace This concept explores how blockchain technology can be leveraged to unlock collaboration and data accessibility for AI development through decentralized marketplaces**.** A blockchain-based marketplace offers a transparent, trust less, and secure environment for data exchange**,** model sharing, and collaborative AI development [12].

2.3 Outcome of the literature review

AI model performance is significantly enhanced by the volume and quality of training data, as evidenced by projects like ImageNet and studies on image enhancement. However, traditional centralized data curation is resource-intensive, highlighting the need for collaborative and distributed data pipelines. While blockchain-based platforms like Ocean Protocol and Singularity NET offer decentralized data exchange, they face challenges including scalability, high energy consumption, and difficulties in ensuring data quality, with incentive systems often prioritizing quantity over utility. Therefore, there's a critical need for improved auditability, decentralized governance, and fair contributor recognition in these systems.

Despite their potential, current decentralized systems face notable limitations**:**

- ✓ Scalability constraints and network congestion
- ✓ High energy consumption, especially on Proof-of-Work-based chains
- ✓ Insufficient mechanisms for ensuring data quality
- ✓ Incentive structures that often favour data volume over utility and fairness

These issues emphasize a pressing need for improved auditability, robust decentralized governance models**,** and equitable contributor recognition systems. Future frameworks must integrate fairness-aware mechanisms that reward data utility and diversity, while ensuring transparency, ethical sourcing, and reliable validation processes.

However**,** traditional centralized data curation methods are both resource-intensive and scalability-limited, prompting a shift toward collaborative and distributed data pipelines. In this context, blockchain-based platform**s** such as Ocean Protocol and SingularityNET emerge as promising solutions for decentralized data exchange and model sharing. These platforms enable data ownership, transparency, and peer-to-peer transactions.

CHAPTER 3

Applications and Requirements

3.1 Applications

1. **Fair Data Collection for AI/ML Training**

   - Ensures diverse and unbiased datasets for training AI models (e.g., facial recognition, voice assistants).
   - Maintains contributor rights and provides traceability for data origins.

2. **Medical and Healthcare Data Crowdsourcing**

   - Collects anonymized patient data securely for disease research and health analytics.
   - Protects privacy while enabling traceable, consent-based data sharing.

3. **Language and Speech Dataset Collection**

   - Gathers regional language or dialect data for natural language processing (NLP).
   - Ensures fair contribution and rewards for underrepresented communities.

4. **Autonomous Vehicle Training**

   - Collects road images, traffic signs, and user-labelled video data for self-driving algorithms.
   - Provides a trusted, verifiable source of crowd-labelled data.

5. **Social and Behavioural Research**

   - Ensures ethical data collection from individuals for psychology, education, and human behaviour studies.
   - Gives participants control and visibility over how their data is used.

3.2 Requirements

Hardware Requirements

Component	**Minimum Requirements**	**Recommended**
Processor (CPU)	Intel i5 or AMD Ryzen 5	Intel i7/Ryzen 7 or higher
RAM	8 GB	16 GB or higher
Storage	250 GB HDD or SSD	512 GB SSD (for faster performance)
Network	Broadband Internet	High-speed Ethernet/Wi-Fi

Software Requirements

- ✓ Operating System
  - Windows 10/11
  - Ubuntu Linux 20.04 or later

- ✓ Blockchain Integration
  - Data submission logging
  - Contributor tracking
  - Incentive/reward distribution

- ✓ Fairness Evaluation
  - Fairness libraries (e.g., AIF360, Fair learn)
  - Support for multiple data types (text, image, audio, etc.)

CONCLUSION

A **federated based data pipeline for sourcing or pooling datasets** offers a transformative approach to ensuring **fairness, transparency, and accountability** in distributed learning among different organisations. Such systems can:

- ❖ Track data provenance and consent
- ❖ Ensure demographic diversity and reduce bias
- ❖ Incentivize ethical behaviour through smart contracts
- ❖ Enable community governance via DAOs
- ❖ Support transparent and auditable fairness evaluations

This approach not only enhances the **trustworthiness** of datasets used in AI/ML systems but also empowers contributors from diverse backgrounds, fostering inclusivity in data-driven innovation. As concerns about bias and data ethics continue to grow, play a key role in building responsible, equitable AI ecosystems.

The integration of federated approach into dataset crowdsourcing pipelines represents a significant advancement in how we approach data fairness, accountability, and inclusivity**.** As AI systems increasingly rely on large and diverse datasets, the risks of bias, exploitation, and lack of transparency grow. A federated data pipeline addresses these issues by offering a trust less, decentralized, and auditable infrastructure for data collection and validation.

Fairness isn’t just ethical, it's foundational to reliable AI performance, especially in sensitive domains like:

- Healthcare
- Criminal justice
- Finance
- Education
- Hiring & recruitment

REFERENCES

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

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

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

[4] 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.

[5] 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

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

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

[8] Jing Chen, Stony Brook University, Algorand

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

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

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

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