The ETA Protocol
An Incentivized Dataset Distribution System

The ETA Foundation

Abstract

The ETA network is a decentralized, blockchain-based platform designed to facilitate the secure and incentivized sharing of machine learning datasets. By addressing the critical issues of data silos, privacy concerns, lack of incentives, and technical barriers, ETA enables a global marketplace where datasets can be accessed and shared transparently. It introduces a token-based economy to reward data providers for participation, ensuring high data availability. ETA leverages blockchain technology to offer censorship-resistant, scalable, and efficient solutions that serve the needs of researchers, developers, and industries reliant on machine learning.

1. Introduction

Background and Motivation

The field of machine learning has witnessed rapid growth in recent years, driven in part by advances in algorithms, such as the creation of the transformer architecture. However, the main driver of progress in machine learning is the availability of large and high-quality datasets. The most powerful models of today, such as GPT-4 (and its derivatives, e.g. o1) and Claude 3.5, are trained on massive datasets containing billions of examples across various domains. In fact, the increases of intelligence between, for example, GPT-2, GPT-3, and GPT-4 are correlated directly with the size of the datasets they were trained on.

In this context, the sharing of machine learning datasets plays a crucial role in advancing research and applications, especially that the research and techniques behind state-of-the-art models are often published and open-source - in such cases, the main barrier to creating similar open models is the lack of access to the datasets used to train them.

Below, we enumerate some of the key challenges brought about by the current state of data sharing in machine learning:

Data Silos and Accessibility

Organizations and institutions often hold valuable datasets in silos, restricting access due to competitive concerns, regulatory compliance, or lack of infrastructure for secure sharing. This fragmentation leads to duplicated efforts, where multiple entities collect similar data independently, wasting resources and slowing down progress.

For example, OpenAI has not released datasets for any of their models beginning with GPT-3, opting only to describe the dataset contents in vague terms. Similarly, Anthropic describes the Claude 3 and 3.5 datasets[1] as a "proprietary mix of publicly available information, (. . .) as well as non-public data from third parties, data provided by data labeling services and paid contractors, and data we generate internally."

This trend of keeping datasets closed-source is made even more concerning by the fact that the models trained on these datasets are often open-source (or at least publicly available), leading to a situation where the most powerful models are trained on data that is not accessible to the broader research community.

Privacy and Security Concerns

Data privacy regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) impose strict guidelines on how personal data can be stored, processed, and shared. Ensuring compliance while sharing datasets is non-trivial, requiring robust mechanisms to protect sensitive information and prevent unauthorized access.

Proprietary datasets are not auditable by the public, making it difficult to ensure that they do not contain personal information, trade secrets, or other sensitive data. This lack of transparency can lead to a situation where models trained on these datasets have a latent knowledge of sensitive information, which could be exploited by malicious actors.

Quality and Standardization Issues

Even when datasets are available, they may suffer from inconsistent formats, lack of proper documentation, or poor data quality. This inconsistency hampers the ability of researchers and developers to effectively utilize the data, necessitating additional preprocessing efforts.

Without an incentive structure rewarding those who share high-quality datasets, there is little motivation for data providers to invest in data curation, cleaning, and documentation. This results in a situation where the quality of public datasets is often subpar, leading to models that are less accurate and more prone to biases.

[1] https://www-cdn.anthropic.com/f2986af8d052f26236f6251da62d16172cfabd6e/claude-3-model-card.pdf

Vulnerabilities of Centralized Systems

The absence of a unified platform that facilitates secure, efficient, and incentivized data sharing further exacerbates these challenges. Traditional centralized systems are vulnerable to single points of failure, censorship, and do not inherently encourage collaborative contributions.

There have been many examples of copyright trolls issuing takedown notices to researchers who have shared datasets that they have collected, leading to a chilling effect on the sharing of datasets. This is particularly concerning to small-scale researchers who may not have the resources to fight such takedown notices or may not have dataset replication strategies in place, making them vulnerable to losing access to the data they need for their research.

Overview of the ETA Protocol

ETA emerges as a solution to these pressing issues by providing a decentralized platform for sharing machine learning datasets. With blockchain technology, ETA ensures transparency, security, and immutability of data within the network. At the same time, it introduces a token-based economy to incentivize data providers, encouraging them to share high-quality datasets while maintaining control over their data.

By addressing these challenges, the ETA Protocol aims to accelerate innovation in machine learning by making high-quality datasets more accessible while respecting privacy and incentivizing participation.

Further in this document we will explain the architecture of the ETA network, the mechanisms for data sharing, the incentive models, and the security and privacy considerations that underpin the protocol.

2. Architecture of the ETA Network

System Components

The ETA Network architecture is designed around three key components: nodes, data storage mechanisms, and a security infrastructure. These components collaborate to ensure seamless, secure, and incentivized data exchanges.

Nodes

There are two primary types of nodes in the ETA Network:

- **Data Providers**: These nodes contribute datasets to the network. Data providers may include research institutions, private companies, or individual contributors. They upload datasets, manage their access permissions, and receive compensation in $ETAI tokens when their data is accessed.

- **Data Consumers**: These nodes are users who access and utilize datasets for machine learning tasks. Consumers can range from academic researchers to industry professionals. They pay for access using $ETAI tokens, thus incentivizing data providers.

Both types of nodes can participate in network governance and decision-making processes.

Data Storage Mechanisms

Datasets in the ETA Network are stored using a hybrid of on-chain and off-chain storage solutions:

- **On-chain storage** is used for metadata, transaction records, and access control lists (ACLs). The blockchain ledger ensures transparency and immutability for all dataset-related transactions.
- **Off-chain storage** is used for the actual dataset files. Due to the large size of machine learning datasets, it is impractical to store them directly on the blockchain. Instead, the network references the datasets using cryptographic hashes stored on-chain, while the datasets themselves are stored in encrypted and authenticated chunks on the Data Providers' nodes.

Security Infrastructure

Security in the ETA Network is paramount. It includes:

- **Encryption**: All data transactions, including dataset submissions and access requests, are secured through encryption. The data encryption keys are stored and managed by the Data Providers, ensuring that only authorized users can access the datasets. Most importantly, even multi-hop connections between nodes are encrypted, ensuring even malicious nodes cannot eavesdrop on the data being transmitted.
- **Access Control**: Data Providers define specific access permissions, ensuring that only authorized users can access their datasets. Role-based access control (RBAC) and public/private key encryption are used to manage dataset access securely. Through gossip protocols, nodes can propagate access permissions across the network, ensuring a correlation between trust and access levels.
- **Authentication**: The network employs a zero-knowledge proof mechanism to authenticate users without revealing sensitive identity information, ensuring privacy and compliance with data protection regulations.

Network Topology

Node Interaction

Nodes in the ETA Network interact through a decentralized, peer-to-peer communication model. Each node maintains a local copy of the blockchain, which records all data transactions and access permissions. When a data consumer requests a dataset, the request is propagated across the network, and the data provider's node verifies the transaction. Once verified, the RBAC system grants access to the dataset, which the consumer can then retrieve from a set of distributed storage nodes.

Scalability Considerations

Large-scale machine learning datasets are large and require efficient storage and retrieval mechanisms, especially when shared across a decentralized network. To alleviate scalability concerns, especially when a high number of requests are made simultaneously, the ETA Network may employ the following strategies:

- **Sharding**: The blockchain is divided into multiple shards, each capable of processing different sets of transactions in parallel, significantly improving throughput. This sharding mechanism could work in conjunction with the data storage mechanism, ensuring that different shards handle different datasets, thus reducing contention, and allowing node operators to specialize in certain types of datasets.
- **Layer 2 Solutions**: By employing off-chain solutions like state channels, the network can execute micro-transactions (e.g., dataset access requests) off-chain, reducing the computational burden on the main blockchain.
- **Distributed Storage Networks**: As datasets grow in size, distributed storage systems like IPFS and decentralized cloud solutions allow for scalable and cost-effective storage, reducing the load on the blockchain itself. The ETA Protocol could integrate with those systems as backup solutions for the off-chain storage of datasets. This way, if a Data Provider goes offline, the dataset can still be accessed by the network.

3. Data Sharing Mechanisms

Dataset Submission

Registration Process

Data providers must go through a registration process to submit their datasets to the ETA network. This process includes:

1. **Trust Verification**: Providers must attain a certain level of trust by staking $ETAI tokens and showing a history of positive interactions within the network. This trust score is used to determine the provider's reputation

and the access permissions they can set for their datasets. Providers with high enough trust scores will be able to submit new datasets to the network.

2. **Dataset Registration**: Providers upload their datasets to an off-chain storage solution while registering the corresponding metadata (e.g., dataset description, size, format, and tags) on the blockchain. This metadata is essential for enabling discovery by Data Consumers.

3. **Access Control Definition**: Providers define initial access permissions for their datasets, specifying who can access them and under what conditions. Permissions can be set based on factors such as user roles, reputation scores, or payment in $ETAI tokens.

Metadata Standards

To ensure interoperability and data consistency, ETA enforces strict metadata standards. These include:

- **Dataset Description**: Providers must include detailed descriptions of their datasets, covering aspects such as data sources, collection methods, and intended use cases.

- **Data Formats**: ETA supports standard machine learning formats such as CSV, JSON, and TFRecord, ensuring compatibility with popular frameworks like TensorFlow, PyTorch, and scikit-learn.

- **Quality Indicators**: Datasets must include indicators of data quality, such as the percentage of missing values, sampling methods, and preprocessing steps.

The ETA Protocol may also include a governance mechanism for defining and updating metadata standards, ensuring that datasets are well-documented and easily discoverable. This mechanism could also be used to punish Data Providers who provide inaccurate or misleading metadata, ensuring that the network maintains a high standard of data quality.

Data Access Protocols

Data Retrieval

Data consumers access datasets by sending an access request to the network. This request includes:

1. **Token Payment**: Consumers pay a fee in $ETAI tokens, which is transferred to the Data Provider upon successful completion of the transaction.

2. **Authentication**: The consumer must authenticate themselves, ensuring they meet the access conditions set by the provider.

3. **Dataset Delivery**: Once authenticated, the original Data Provider prepares a list of nodes that store the dataset chunks. The consumer then

retrieves the dataset chunks from these nodes and reconstructs the complete dataset locally.

4. **Decryption**: Individual dataset chunks are decrypted using the encryption keys provided by the Data Provider, ensuring that only authorized users can access the data.

Access Controls

Data providers can set detailed access controls using role-based access control (RBAC) mechanisms. Permissions can be set based on:

- **User Role**: For example, academic users may receive discounted or free access to certain datasets, while commercial entities are required to pay a higher fee. Those roles can be defined by the Data Provider or by the network governance.
- **Reputation Score**: The ETA Network incorporates a trust system, where users with high scores (based on previous interactions) may gain faster or prioritized access to datasets.
- **Token Balance**: Access to certain datasets may require users to hold a minimum balance of $ETAI tokens in their wallets.

4. Incentive Models

Economic Framework

Token Economics

$ETAI tokens serve as the primary medium of exchange within the network. The tokens are used for:

- **Dataset Access**: Consumers pay in $ETAI tokens to access datasets.
- **Transaction Fees**: Small transaction fees, denominated in $ETAI tokens, are incurred during dataset exchanges, supporting network maintenance and rewarding node operators.
- **Staking and Trust**: Data providers stake $ETAI tokens to establish trust and reputation within the network. Staked tokens can be slashed in case of malicious behavior, creating a strong incentive to act in good faith.

The token economy must be designed to encourage both the sharing and consumption of high-quality datasets, driving value creation in the ecosystem. This can be achieved through mechanisms such as token burns, where a portion of the transaction fees is used to buy back and burn $ETAI tokens, reducing the overall token supply and increasing the value of the remaining tokens.

Incentive Structures

The incentive structures of the ETA network revolve around rewarding participants who contribute valuable data or enhance existing datasets. Key mechanisms include:

- **Data Providers**: Receive $ETAI tokens whenever their datasets are accessed or utilized by consumers.
- **Data Consumers**: In some cases, consumers who contribute meaningful insights (e.g., annotations or improvements) to datasets may be rewarded through the network's curation system. Similarly, consumers who provide misleading or incorrect information may be penalized through token slashing.

Reward Distribution

Data Providers

Data providers earn tokens whenever their datasets are accessed or referenced. The more frequently a dataset is used, the higher the rewards. A portion of the transaction fees also goes to the provider as a continuous revenue stream, providing long-term incentives for sharing. The specific trust score of the provider will also influence the rewards they receive, with higher trust scores leading to higher rewards.

Data Consumers

Consumers may participate in validating the quality and integrity of datasets. They may propose improvements, such as fixing missing data or standardizing formats. Curators receive a share of the tokens whenever their enhancements are adopted by the network or contribute to improving the dataset's utility.

5. Node Communication Protocols

Communication Models

Data Propagation

Information in the ETA Network is propagated through a gossip protocol, where nodes share updates (e.g., new datasets, transactions) with neighboring nodes, and this information gradually spreads throughout the network. This decentralized propagation model ensures that all nodes remain up-to-date without relying on a central authority.

Message Formatting

Node communication relies on well-defined message structures, which ensure that datasets, transactions, and metadata are exchanged in a standardized manner.

Each message follows a format consisting of:

- **Header**: Contains metadata about the message, such as timestamp, message type, and node ID.
- **Payload**: Includes the actual data being exchanged, such as transaction details or dataset metadata.
- **Signature**: Ensures message authenticity, using digital signatures that allow recipient nodes to verify the sender's identity.

Synchronization Methods

State Consistency

To maintain a consistent state across the network, ETA uses a consensus mechanism based on Byzantine Fault Tolerance (BFT). This algorithm ensures that all honest nodes agree on the order and content of transactions, preventing discrepancies and double-spending.

Latency Reduction

To minimize latency in communication, ETA implements several optimizations:

- **Caching**: Frequently accessed datasets or metadata are cached at the edge nodes, reducing retrieval time for popular datasets.
- **Batch Processing**: Transactions are processed in batches, reducing the overhead associated with individual transactions and improving throughput.

6. Security and Privacy Considerations

With extensive data sharing and financial transactions occurring within the ETA network, robust security and privacy measures are essential to protect user data and prevent malicious activities. It must be noted that ETA's censorship resistance mechanisms must be structured in such a way that they do not enable the sharing of morally reprehensible content.

ETA aims to protect user data through encryption, zero-knowledge proofs, and secure deletion protocols, ensuring that data remains confidential and tamperproof. Similarly to IPFS, ETA users may maintain censorship lists, which would allow them to blacklist certain datasets or nodes that are known to contain harmful or illegal content.

Data Encryption

Encryption is a fundamental component of the ETA network, ensuring that data remains secure and confidential throughout its lifecycle. To ensure fair access to datasets, the network must transmit data in encrypted form, ensuring that only authorized users can access the original content.

Encryption Techniques

All datasets stored off-chain are encrypted using symmetric encryption algorithms, such as AES-256. This ensures that even if unauthorized parties gain access to the storage system, the data remains unintelligible. The encryption process is mathematically represented by the following:

$$C = E_k(M)$$

Where M is the original message (the dataset), E_k is the encryption function using key k, and C is the resulting ciphertext. To ensure the integrity of the data, the encryption process is combined with a hashing function:

$$H(M) = h_k(M)$$

Here, h_k is the keyed hash function, ensuring that any tampering with the dataset can be detected by comparing hash values. All data in transit between nodes is protected using TLS 1.3, providing secure communication channels.

Key Management

Managing encryption keys is crucial to maintaining the confidentiality of datasets. ETA employs a decentralized key management scheme, where each data provider generates their own keys using public-private key cryptography:

$$K_{\text{pub}}, K_{\text{priv}} \quad \text{where} \quad K_{\text{priv}} = \mathcal{D}(K_{\text{pub}})$$

Here, K_{pub} is the public key used for encryption, and K_{priv} is the private key used for decryption. The keys are securely distributed using a threshold-based secret sharing scheme to prevent single points of failure.

Privacy Protocols

Ensuring privacy is an important consideration within the ETA Protocol, particularly when dealing with sensitive data that could fall under regulatory frameworks like GDPR or HIPAA. This section outlines the privacy measures implemented by the network.

Anonymity Measures

User anonymity is preserved using zero-knowledge proofs (ZKP). These cryptographic techniques allow users to prove possession of certain data or permissions without revealing the actual data. Formally, a zero-knowledge proof allows a prover to convince a verifier that a statement is true without providing any additional information beyond the validity of the statement:

$$\text{ZKP}: \quad \mathcal{P}(x) \longrightarrow \mathcal{V}(x): \quad P(x) \Rightarrow \text{TRUE}$$

Where $\mathcal{P}$ is the prover, $\mathcal{V}$ is the verifier, and x is the data that is never exposed. This ensures that user identities and their activities on the network remain anonymous.

Such a proof scheme is particularly useful when combined with ETA's encryption mechanisms, as it allows Data Providers to prove that they have the decryption key for a dataset without revealing the key itself. This can be used to verify that a Data Provider has the necessary permissions to access a dataset without exposing the key to other network participants.

Compliance with Regulations

The ETA network incorporates features designed to comply with stringent data privacy regulations, such as GDPR's "right to be forgotten." This is implemented using secure data deletion protocols, where encryption keys associated with the dataset are destroyed upon request, rendering the data irretrievable.

By ensuring that the decryption key k is no longer available, the original dataset M is effectively erased.

7. Scalability and Performance

Performance Optimization

Given the size and complexity of machine learning datasets, the ETA network incorporates several optimization strategies to ensure high performance and low latency.

Throughput Enhancements

The throughput of the network is increased by leveraging sharding techniques, where the blockchain is partitioned into multiple shards that operate in parallel. This division allows for the concurrent processing of multiple transactions, represented mathematically as:

$$T_{\text{total}} = \sum_{i=1}^{n} T_{\text{shard}_i}$$

Where T_{total} is the overall transaction throughput, and T_{shard_i} is the throughput of the i-th shard. This architecture dramatically increases the number of transactions that can be processed simultaneously.

Load Balancing

To prevent network bottlenecks, load balancing algorithms distribute the workload evenly across available nodes. The load L_i assigned to node i is determined dynamically using:

$$L_i = \frac{\sum_{j=1}^{m} W_j}{n}$$

Where W_j is the weight of the j-th task, and n is the number of available nodes. This ensures that no single node is overwhelmed by excessive requests.

Storage Solutions

Given the size of datasets typically used in machine learning, efficient storage is a necessity. The ETA network combines on-chain and off-chain storage methods to manage large-scale data effectively.

Off-Chain Storage

Datasets are stored off-chain using a bespoke storage solution that ensures high availability and low latency. The off-chain storage system is designed to handle large files efficiently, enabling rapid retrieval and access. The network references these datasets using cryptographic hashes stored on-chain, ensuring data integrity and immutability.

As a backup solution, the ETA network may integrate with decentralized storage solutions like IPFS, which distribute data across multiple nodes.

The advantage of this approach lies in its scalability and redundancy. The off-chain storage of dataset D is referenced on-chain by a unique hash:

$$\text{Stor}(D) = H(D)$$

Where $H(D)$ is the cryptographic hash that acts as the pointer to the dataset. This ensures that the actual dataset size does not burden the blockchain, while the integrity of the data is verifiable through the hash.

Data Sharding

To further improve efficiency, the ETA network implements data sharding, where large datasets are split into smaller chunks that can be stored and retrieved independently:

$$D = \{D_1, D_2, \ldots, D_n\}$$

Each shard D_i is stored on separate nodes, and the original dataset can be reconstructed using a concatenation operation:

$$D = D_1 \oplus D_2 \oplus \cdots \oplus D_n$$

This method allows for parallel retrieval of dataset components, significantly speeding up the access time.

8. Governance and Community Involvement

Decision-Making Processes

The governance of the ETA network is decentralized, enabling participants to influence the network's direction. This process is based on a voting system where stakeholders propose changes to the protocol and vote on their implementation.

Protocol Upgrades

Protocol upgrades follow a formal process in which participants submit proposals to improve network functionality. Each proposal P is subjected to a vote, and the total weight of the votes $V(P)$ is determined by the number of tokens held by each voting participant:

$$V(P) = \sum_{i=1}^{n} T_i \times v_i$$

Where T_i is the number of tokens held by participant i, and v_i is their vote (either 1 for approval or 0 for rejection). A proposal is accepted if the total weighted vote exceeds a predefined threshold $V_{\min}$.

Community Roles

The ETA network fosters a collaborative environment where community members take on specific roles to enhance the ecosystem.

Contributor Recognition

Contributors who actively participate in improving datasets or curating data are recognized through a reputation system. The reputation score R of each participant is updated based on their actions within the network:

$$R(t+1) = R(t) + \Delta R$$

Where ΔR represents the change in reputation score based on positive or negative contributions.

9. Use Cases and Applications

Research and Development

The ETA network has significant potential in academic and research settings, particularly in fostering collaboration across institutions. Researchers can share datasets that are otherwise restricted due to institutional silos or lack of infrastructure. This open-access model accelerates innovation and promotes transparency in data-driven research.

For instance, the use of ETA in collaborative research on natural language processing (NLP) could lead to better language models by enabling access to diverse multilingual datasets. Academic researchers contribute data in exchange for recognition and token rewards, which can then be used to access other datasets for their own research purposes.

Industry Implementation

In the healthcare sector, the ETA network addresses the critical issue of data fragmentation across hospitals and medical institutions. Machine learning models for disease prediction, patient management, and medical image analysis rely on diverse datasets that are often difficult to obtain due to privacy concerns and competitive barriers. ETA's secure, decentralized platform ensures that such datasets can be shared while complying with privacy regulations like HIPAA and GDPR.

In financial modeling, institutions can access proprietary datasets that are normally kept siloed. Machine learning models used for credit scoring, fraud detection, and market prediction benefit from larger datasets, reducing biases and improving accuracy. The incentivization mechanisms in ETA encourage financial firms to share their anonymized datasets, fostering a more collaborative industry ecosystem.

10. Future Work and Roadmap

Upcoming Features

The ETA Protocol's future development includes several key enhancements designed to improve functionality and user experience. One planned feature is the introduction of **smart contracts** that allow automated licensing of datasets, enabling dynamic pricing based on demand. Additionally, the network is exploring **differential privacy** mechanisms that allow data providers to share useful insights from sensitive datasets without revealing the underlying data.

Mathematically, differential privacy ensures that for any two neighboring datasets D_1 and D_2, the probability distribution of the output of a function f remains nearly identical:

$$\Pr[f(D_1) = y] \approx \Pr[f(D_2) = y]$$

This guarantees that individual data points cannot be inferred from the output, providing privacy while still enabling data sharing.

Research Directions

The ETA network's research focus includes the integration of **homomorphic encryption**, which would enable computations to be performed directly on encrypted data without requiring decryption. This has profound implications for secure machine learning, allowing models to be trained on sensitive data while maintaining privacy.

Homomorphic encryption allows operations such as addition and multiplication on ciphertexts:

$$E(x_1 + x_2) = E(x_1) + E(x_2)$$

This allows models to be trained without ever revealing the underlying data, preserving both privacy and functionality.

11. Conclusion

In summary, the ETA network provides a comprehensive solution to the challenges surrounding machine learning dataset sharing. By leveraging blockchain technology, it ensures data security, privacy, and incentivization, creating a collaborative ecosystem that benefits researchers, industries, and developers. As machine learning continues to advance, the ETA network aims to drive innovation by making high-quality datasets more accessible, without compromising privacy and security. This collaborative platform encourages stakeholders to contribute, share, and benefit from the shared knowledge and datasets, participating in the next wave of machine learning advancements.