DataFi: A Decentralized Data Network for AI Training

Unlocking the True Value of Data, The New Oil of the AI Era

Crypto Gurus (Group 7)
November 2025

Gasparini Gabriele, Turiano Lorenzo, Cantillo Pablo, Cinquepalmi Flavia

Contents

Abstract

Artificial Intelligence depends on the quality, diversity, and accessibility of training data. Yet the current data landscape is fragmented and inequitable. High-quality datasets are centralized within a few corporations, locked behind paywalls, and often lack transparency regarding provenance, bias, and consent. Researchers and smaller labs face prohibitive costs and limited access, resulting in models trained on narrow, unrepresentative samples that perpetuate systemic bias and hinder generalization. Moreover data creators, individuals generating valuable real-world information, rarely capture any of the value their data enables. Privacy concerns, ownership ambiguity, and the absence of verifiable data-exchange mechanisms further stall collaboration and innovation.

DataFi addresses these challenges by introducing a blockchain-based decentralized data marketplace that enables secure sharing, verification, and monetization of datasets. Through federated learning, a technique that trains machine learning models collaboratively across many participants without transferring their actual data, contributors can participate without exposing raw information, while on-chain provenance ensures authenticity, traceability, and fair compensation. Its native token, $DATA, powers payments, staking, governance, and marketplace liquidity, aligning economic incentives across all participants. Beyond technical efficiency, *DataFi* promotes a new digital profession, the *dataset entrepreneur*, empowering anyone to earn income by creating, curating, and selling datasets used for AI training. By democratizing access, incentivizing quality, and embedding transparency into the data economy, *DataFi* aims to transform the foundation on which AI is built.

1 Executive Summary

The adoption of AI is accelerating across industries, but model performance is constrained by data scarcity, opacity of provenance, and uneven access. Meanwhile, regulations such as the EU AI Act [1] demand traceability and accountability in training pipelines. *DataFi* solves these issues by combining (i) **blockchain** for immutable provenance and automated licensing; (ii) **federated learning** so models can learn from distributed datasets without exposing raw data; and (iii) **tokenized incentives** so creators and validators are fairly rewarded.

Investment thesis:

- Strong macro tailwinds (AI scale-up, governance, and compliance).
- Clear market need for compliant, auditable, bias-aware datasets.
- Tokenized incentives unlock supply (data creators) and demand (AI developers).
- Unique social upside: job creation for *data entrepreneurs*.

2 Problem and Opportunity

2.1 Industry Focus

This business opportunity targets the **artificial intelligence and data infrastructure industry**, a rapidly expanding sector underpinning automation, analytics, and generative AI applications across every field from healthcare and finance to manufacturing and education. The core bottleneck in this industry lies not in model architecture, but in the **availability, quality, and compliance of training data**. As AI becomes mainstream, data has emerged as the primary constraint on scalability, transparency, and fairness.

2.2 The Data Bottleneck

High-quality datasets remain difficult to access because they are expensive, fragmented, and often locked within the proprietary systems of large technology firms. For instance, Zhu et al. [2] describe Microsoft's internal data infrastructure as an *exabyte-scale* system, one of the largest in the world, spanning over 300,000 machines and handling billions of processing tasks each day. Likewise, Amazon's Redshift Spectrum platform supports *exabyte-scale, in-place queries* across its S3 data storage environment [3]. These examples illustrate how data accumulation has become **highly centralized**: big companies possess both the scale and the infrastructure to collect, clean, and analyze massive datasets internally. In contrast, smaller AI labs, startups, and research teams lack the resources and permissions to access equivalent data. They must instead spend significant time and money sourcing, validating, and legally clearing smaller datasets, often of lower quality and diversity, making it extremely difficult to compete or comply with emerging regulatory requirements for transparency and provenance.

2.3 Market Context

According to multiple industry reports, the global AI training data market is projected to exceed **$10 billion by 2030**, growing at a double-digit CAGR [4]. Rising regulatory pressure driven by the EU AI Act [1] and emerging data governance frameworks, requires greater traceability, fairness, and documentation in training pipelines. Buyers increasingly seek **bias-aware, compliant, and well-documented datasets**, while suppliers want predictable monetization and intellectual property protection.

2.4 Opportunity

A major opportunity exists to create a **decentralized and compliant data marketplace** where ownership is provable, usage is traceable, and compensation is automatic. Such a

platform would align incentives so that more and better data flows into AI development, legally, transparently and equitably.

2.5 Social Impact: Job Creation

DataFi transforms data work into a first-class digital profession. **Data creators** such as students, freelancers, labs and SMEs can earn recurring income by producing domain-specific, bias-aware, and well-documented datasets. This democratizes participation in the AI economy and supports inclusive global innovation.

3 Stakeholders and End-Users

3.1 End-Users of the DataFi Network

The primary end-users are the individuals and organizations that directly benefit from accessing, consuming, or contributing to the *DataFi* platform. They are the ones who use the network to achieve practical outcomes, such as training AI models, sourcing verified datasets, or earning through data creation.

- **AI Developers and Startups:** Companies, research labs, and individual developers who require high-quality, bias-aware, and traceable datasets for training, fine-tuning, and validating AI models. These users access *DataFi*'s marketplace to obtain verified datasets with transparent licensing and provenance information, reducing data acquisition costs and accelerating model development.
- **Data Creators:** Individuals, freelancers, students, and small teams who create, clean, label, and structure datasets for sale. They use the platform to publish their work, gain visibility, and earn $DATA tokens each time their datasets are purchased or used in model training.
- **Data Consumers in Enterprises:** Corporate data science teams and innovation units that need specialized or domain-specific datasets (e.g., healthcare, finance, mobility). They use *DataFi* to acquire compliant and auditable data assets for in-house AI projects without building costly proprietary pipelines.

3.2 Stakeholders Involved in the Solution

Stakeholders are the broader ecosystem participants who influence, govern, or benefit indirectly from DataFi's operation. They help maintain integrity, scalability, and compliance across the platform.

- **Data Providers (Institutions and Enterprises):** Organizations that already own proprietary datasets. They contribute data under controlled licensing conditions and receive financial rewards for usage, while maintaining ownership and intellectual property rights.
- **Data Validators and Curators:** Accredited professionals or community members who verify dataset quality, structure, and compliance. They play a governance role by ensuring only valid, bias-assessed, and well-documented datasets are published. Validators earn $DATA rewards for maintaining the integrity of the marketplace.
- **Regulators and Compliance Authorities:** National and regional agencies (e.g., under the EU AI Act [1]) overseeing AI transparency and data ethics. They interact with DataFi through audit interfaces that provide immutable provenance trails and data governance reports.
- **DataFi DAO and Token Holders:** The decentralized governance body managing system parameters, reward mechanisms, and policy updates. Token holders act as strategic stakeholders, influencing the long-term direction of the ecosystem.

3.3 Motivations and Value Alignment

- **For End-Users:** Access to affordable, bias-aware datasets; opportunities to monetize data creation work; improved efficiency in AI development.
- **For Stakeholders:** Transparent governance, new revenue models, and a framework that supports compliance, auditability, and trust in AI data pipelines.

4 The DataFi Solution

4.1 Core Capabilities

At a high level, the *DataFi* network operates as a decentralized ecosystem connecting data providers, validators, and AI developers through blockchain-based smart contracts.

1. **On-chain Provenance:** each dataset is fingerprinted (cryptographic hash) and registered on-chain with metadata, terms, and lineage.
2. **Automated Licensing:** smart contracts enforce access rules, pricing, and revenue sharing; usage events trigger payouts.
3. **Privacy-Preserving Training:** *DataFi* integrates **federated learning** and secure aggregation to enable AI models to learn collaboratively from distributed datasets without exposing raw data. Federated learning is a method where the training process

happens directly on each data provider's device or server: the model learns from the data locally, and only the *updated model's weights* (not the data itself) are shared back to the end-users [5]. This approach ensures data confidentiality, reduces the risk of breaches, and complies with privacy regulations such as GDPR [6].

4. **Bias & Quality Tooling:** curation templates, datasheets [7], and bias scores help buyers evaluate suitability and risk.

5. **Decentralized Storage:** *DataFi* uses the IPFS [8] for secure and distributed data storage, a decentralized way to store and share files on the internet. Instead of relying on a single server (like Google Drive or AWS), IPFS spreads files across many independent nodes all over the world. All datasets are encrypted before being stored, and access is managed through **license keys** and controlled **off-chain gateways**, allowing only authorized users to retrieve and decrypt the data.

4.2 Core Workflow

1. **Dataset Listing and Registration:** Data providers register their datasets on the *DataFi* platform. Each dataset is encrypted, stored on IPFS, and receives a unique on-chain fingerprint (*hash*) recorded via a smart contract. The provider specifies access rules, licensing terms, and whether the dataset can be used for direct download or only through federated learning.

2. **Data Validation and Quality Control:** Validators stake $DATA tokens to participate in dataset verification. They inspect metadata, documentation, and dataset structure to ensure compliance with DataFi's quality and governance standards. Only validated datasets become visible to buyers. Validators earn rewards for maintaining data integrity.

3. **Data Access and Usage:** Buyers (AI developers or enterprises) interact with datasets depending on the provider's chosen mode:

   - **Raw Dataset Access:** The buyer pays the provider in $DATA tokens to obtain access rights. After payment, the smart contract issues a decryption key through an off-chain gateway, allowing the buyer to download and freely use the raw dataset for AI training or analytics.
   - **Federated Learning Access:** If the provider opts for privacy-preserving sharing, the raw dataset never leaves their infrastructure. Instead, the buyer uploads their training code or model through a secure interface (e.g., Google Colab or Jupyter Notebook integration). The provider's node executes the training job locally using the dataset and returns only the **updated model weights or gradients** to the

buyer. This ensures that the model learns from the data while the provider retains full control and confidentiality over their dataset.

4. **Economic Settlement and Governance:** All transactions such as dataset purchases, validator rewards, federated training fees are settled automatically through $DATA smart contracts deployed on an Ethereum Layer 2 network (e.g., Polygon or Arbitrum). Gas fees are paid in ETH, while internal incentives and governance functions use $DATA. Governance decisions (e.g., protocol updates, staking parameters, and dataset category priorities) are handled through the **DataFi DAO**, where token holders can vote on proposals.

Technical Summary: This architecture ensures that:

- Data remains **secure and private**, never leaving the provider's control.
- Transactions and licensing are **transparent and automated** via blockchain.
- Model training can occur in a **privacy-preserving** manner using federated learning.
- The ecosystem remains **self-sustaining and trustless**, driven by token-based incentives.

4.3 Protocol Choices

DataFi is built as a **Proprietary Blockchain-Based Software Solution model** [9] (Business Model Continuum). This means we operate on decentralized blockchain infrastructure (Ethereum Layer 2) but develop our own proprietary platform and software layer on top of it. In simpler terms, Ethereum provides the trusted, decentralized backbone, while *DataFi* builds the specialized data marketplace and governance system that unlocks real business value.

4.4 Ethereum as the Network Gasoline

A core design decision for *DataFi* is to operate on the Ethereum ecosystem, using **Ether (ETH)** as the underlying "gasoline" to pay for network transactions, while maintaining the $DATA token as the internal economic and governance unit.

Rationale and Benefits

- **Security and Reliability:** Ethereum provides the most battle-tested and secure smart contract environment in the blockchain space, supported by thousands of validators and institutional-grade infrastructure. This ensures long-term trust and technical robustness for a project that manages sensitive and high-value datasets.

- **Interoperability:** By aligning with the Ethereum ecosystem, *DataFi* gains instant compatibility with existing DeFi protocols, custodial wallets, exchanges, and decentralized identity systems. This enhances user adoption and credibility with investors and enterprise partners.
- **Layer 2 Efficiency:** To mitigate high gas fees, *DataFi* will deploy on an Ethereum Layer 2 network such as *Polygon* [10] or *Arbitrum* [11]. These layers retain Ethereum security while offering faster transactions and drastically lower costs.
- **Economic Clarity:** Using ETH for gas decouples transaction costs from the $DATA economy. The $DATA is used exclusively for staking, governance, payments, and marketplace activity, preventing inflationary pressure and maintaining a clear token utility model.
- **User Experience:** For users without ETH, the platform can implement meta-transactions or *gas relayers* to allow gas fees to be paid indirectly in $DATA, offering a seamless on-boarding experience. This is a new standard for Ethereum called **ERC-4337** (**"account abstraction"**) which allows for "paymaster" contracts

Strategic Value This hybrid model (ETH for gas, $DATA for internal incentives) combines the **trust and stability of Ethereum** with the **flexibility of DataFi's native economy**. It positions the project within the most recognized blockchain ecosystem while ensuring scalability and cost efficiency through Layer 2 deployment.

5 Organizational Perspective & Governance

5.1 Operating Model

- **AI Engineering**: federated learning SDK, quality/bias scoring.
- **Blockchain Development**: licensing, payments, staking, slashing.
- **Data Governance**: validation criteria, documentation standards, ethical review.
- **Legal/Compliance**: GDPR, data sharing agreements, IP/licensing frameworks.

These four divisions form a cohesive operating model: AI engineers guarantee performance and fairness; blockchain developers ensure transparency and automation; data governance specialists safeguard integrity and ethics; and legal experts maintain compliance and trust. This configuration was chosen deliberately to make *DataFi* technically resilient, ethically sound, and legally robust,a prerequisite for scaling in the highly regulated AI data economy.

5.2 DataFi DAO

Token holders propose and vote on fees, reward curves, validator criteria, dataset category priorities, and grants. Public dashboards provide transparency on inflows/outflows, validator performance, and treasury data.

Governance draws on a meritocracy, consensus-based model observed in open-source projects like Bitcoin [12] and Ethereum [13].

Governance tiers include:

- **Validators (producers)**: maintain service reliability through staking and slashing.
- **Developers (providers)**: propose upgrades and ensure technical security.
- **Token holders (appropriators)**: vote on DAO proposals, promoting community accountability.

This structure supports sustainability through shared accountability and incremental adaptation, two of the key "polycentric practices" from the Bitcoin governance model.

6 Business Perspective

6.1 Product & Services

- **Marketplace**: list, discover, license datasets; bias/quality indicators; provenance explorer.
- **APIs/SDKs**: integrate data access and federated learning into MLOps.
- **Analytics**: demand signals, pricing intelligence, dataset performance, and audit logs.
- **Compliance Suite**: enterprise SSO, KYC/AML where applicable, policy enforcement, and exportable audit trails.

From the business model standpoint, this falls under the *application layer* of the blockchain stack (Software–Protocol–Infrastructure). The marketplace functions as a decentralized timestamping and identity certification service (two of the primary blockchain business applications) ensuring immutability and authenticity of digital assets in the AI ecosystem.

6.2 MVP Scope

Launch with text/image datasets suited for LLM and vision fine-tuning; implement core licensing contracts, staking for creators/validators, and basic federated training pilots with academic/industry partners.

This MVP approach mirrors incremental development strategies in open blockchain environments, where early deployment validates user demand and governance parameters before scaling.

6.3 Revenue Model

- **Transaction fee** ($\sim$3%) on every transaction made in the marketplace. This applies whether the buyer purchases a raw dataset or pays to run a federated learning job.
- **Premium tiers** for large organizations that can subscribe to provide advanced features.
- **Advertising and Featured Dataset Placement**. Data providers can pay to highlight or feature their datasets on the homepage or search results for greater visibility.
- **Token Utility and Staking Revenue**. Users must stake $DATA tokens to participate as validators or premium creators. A portion of staking yields or transaction fees can be redirected to the DAO treasury (0.5%), providing a recurring income source for the platform.

According to the Business Model Continuum, revenues are derived at the software layer through access and usage fees rather than infrastructure control, enabling scalability without the costs of fully proprietary ecosystems.

6.4 Risks & Mitigations

- **Data Quality Variance**: address via staking/slashing, validator reputation, and bias/quality scoring.
- **Regulatory Complexity**: strict licensing, consent management, and localized policy modules.
- **Token Volatility**: fee options in stablecoins with $DATA discounts; treasury risk controls.

Consistent with blockchain business considerations, key operational risks include outsourced accountability (public-network infrastructure) and market adoption barriers. These are mitigated through internal governance mechanisms, transparent dashboards, and DAO-managed funding reserves, reflecting best practices from decentralized application management.

7 Token Economics: The $DATA Utility

7.1 Utilities

The design of the $DATA token serves as the primary currency used across the platform. It aligns incentives across creators and AI developers, enabling a scalable data marketplace.

1. **Payment Medium (Datasets/Services)**: developers rely on $DATA to purchase dataset licenses, run federated learning jobs, and access pre-trained models, scripts, and plugins. As it becomes the payment medium for all core services, the platform creates continuous demand for the token.

2. **Staking for Creators/Validators**: staking $DATA acts as reputational collateral to maintain marketplace integrity. Creators must stake tokens to publish premium datasets, while validators stake to participate in quality assurance processes. This mechanism ensures reliable contributor participation and discourages misconduct.

3. **Access to Premium Features**: holding or staking $DATA provides access to premium tools improving dataset visibility and developer productivity. Higher staking tiers unlock dashboards with detailed analytics, performance insights, market trends, priority listing placement, and faster federated training queueing. These incentives reinforce long-term ecosystem engagement.

4. **Liquidity & Staking Rewards**: users contributing $DATA to liquidity pools on decentralized exchanges may earn additional yield while strengthening market stability. $DATA may also be staked natively within the protocol, granting participants a share of the fees generated by marketplace activities. This dual-reward system deepens liquidity and aligns token holders with ecosystem growth.

5. **Discounts & Incentives**: payments made with $DATA unlock exclusive benefits through partnerships with external companies. Users who pay for third-party tools, products, or services using the $DATA token may receive special discounts and preferential pricing as part of these collaborations.

7.2 Economic Impact

DataFi establishes a new global profession: the *data entrepreneur*. Creators can produce valuable datasets, ranging from medical imaging annotations and ESG corpora to cybersecurity logs or geospatial time-series, accompanied by standardized documentation, metadata, and bias assessments. Once published, these datasets generate recurring income with each license purchased.

Since the marketplace model is digital, permissionless, and globally accessible, anyone with domain expertise can participate. This creates a scalable economic ecosystem in which contributors are continuously rewarded as their datasets fuel real AI workloads across industries. *DataFi* effectively transforms high-quality data creation into a global economic opportunity.

8 Roadmap and Implementation Plan

8.1 Phased Plan

The implementation strategy balances technical development, compliance, and early market adoption.

- **Phase 1 (0–3 months)**: core contracts (registration, payments, staking), marketplace alpha, pilot creators.
- **Phase 2 (4–8 months)**: federated learning integration, institutional onboarding, validator program, DAO bootstrapping.
- **Phase 3 (9–18 months)**: full DAO activation, vertical expansion (healthcare, finance, geospatial), enterprise compliance suite.

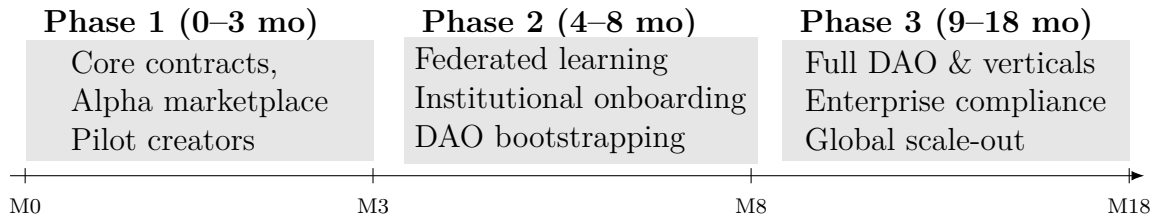

Figure 1: Implementation timeline with major milestones.

9 Financials and Funding

9.1 Seed Round Request

- **Amount:** $1.0M
- **Allocation:** 40% platform & blockchain development; 30% compliance & security; 20% marketing/partnerships; 10% DAO treasury/liquidity.

9.2 Use of Proceeds

Prioritize hiring (smart contracts, federated learning), audits, enterprise pilots, and strategic partnerships (vertical datasets).

10 Why Invest in DataFi

The current landscape makes *DataFi* a high-leverage investment opportunity offering superior solutions. The convergence of AI adoption, data governance requirements, and decentralization trends opens a window for a compliant, privacy-preserving data marketplace.

- **Regulatory Pull**: increasing provenance and auditability requirements make compliant, well-documented datasets essential for modern AI-driven organizations.
- **Clear Differentiation**: *DataFi* focuses on data-centric job creation and privacy-preserving training via federated learning, two gaps largely neglected by existing platforms.
- **Strong Network Effects**: more creators produce richer datasets, attracting more buyers; increased activity generates higher rewards, reinforcing continuous ecosystem growth.
- **Scalable Architecture**: the platform's modular, chain-agnostic design and native MLOps compatibility support enterprise integration and long-term scalability.

11 Conclusion

As high-quality, transparent, and compliant data becomes the defining competitive advantage in AI, *DataFi* offers a powerful solution tailored to this reality. By combining blockchain-based provenance with privacy-preserving federated learning, the platform ensures that organizations can access trustworthy datasets while maintaining strong privacy and regulatory alignment. *DataFi* leverages the growing importance of data by creating a marketplace where ownership is verifiable, usage is traceable, and incentives are aligned. Its token-driven economy rewards creators, validators, and contributors, turning data production and curation into a viable economic opportunity. In a landscape where AI regulation is tightening and demand for bias-aware, auditable datasets continues to accelerate, *DataFi* positions itself as essential infrastructure for responsible and scalable AI development. By making data more accessible, secure, and economically fair, it strengthens the foundations on which the next generation of AI systems will be built.

References

[1] European Union. Artificial intelligence act (regulation (eu) 2024/1689). Official Journal of the European Union, 2024. Regulation of the European Parliament and of the Council.

[2] Yiwen Zhu, Subru Krishnan, Konstantinos Karanasos, Isha Tarte, Conor Power, Abhishek Modi, Manoj Kumar, Deli Zhang, Kartheek Muthyala, Nick Jurgens, Sarvesh Sakalanaga, Sudhir Darbha, Minu Iyer, Ankit Agarwal, and Carlo Curino. Kea: Tuning an exabyte-scale data infrastructure. 2021.

[3] Amazon Web Services. Amazon redshift spectrum: Exabyte-scale in-place queries of s3 data, 2023. Accessed November 2025.

[4] MarketsandMarkets Research Pvt. Ltd. Ai training dataset market size, trends, and industry forecast (2030), 2024. Estimated market size USD 2.82 billion in 2024, projected USD 9.58 billion by 2029 at a CAGR of 27.7%.

[5] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 2021.

[6] European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). Official Journal of the European Union, 2016.

[7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets.

[8] Juan Benet. IPFS – Content Addressed, Versioned, P2P File System. White paper, 2014.

[9] Università Bocconi. Blockchain and Cryptoasset: The Business Perspective (S20), 2025. Lecture slides.

[10] Polygon:. Ethereum's internet of blockchains. White paper, 2021.

[11] Harry Kalodner, Steven Goldfeder, Xueyuan Chen, Seth Weinberg, and Edward Felten. Arbitrum: Scalable, private smart contracts. White paper, 2021.

[12] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. White paper, 2008.

[13] Gavin Wood. Ethereum: Yellow paper, 2023. Originally published in 2014. Latest revision.