OpenSet

Accelerating Innovation through Decentralized Data Infrastructure

Jake Zegil
Andrew Toews

December 1st, 2021

Abstract

"Data is the new gold." - Albert Einstein

A new age of internet companies, steeped in trillions of dollars of data sits at the top of the S&P 500. Revenues are no longer a function of who has the best technology; rather, who has the best data? Data informs strategic decision making, structures our conversations and social discovery mechanisms, and trains the artificial intelligence that is ever more present in our workplaces and homes.

The internet generates over a trillion (1,000,000,000,000) MB of data per day, and most of that data sits in private databases. Data is not shared or improved, and corporations employ exploitative (as opposed to collaborative) data collection practices to maintain data moats, which are cheaper to build and maintain than a technical advantage.

In this paper we propose an alternative economic model - a data storage and ETL ecosystem (OpenSet) that rewards all stakeholders by enabling permissionless participation in data creation, ownership, and financialization.

1 Introduction

Businesses have engaged in anti-competitive practices since the dawn of trade. Negating a competitor's ability to operate has historically acted as a powerful centralizing force in global commerce, only meekly opposed by regulation. Recently, however, innovations in decentralized protocols and Web3 have created the tools and infrastructure required for smaller agents to collaborate with high degrees of trust and coordination, effectively out-innovating and out-competing centralized organizations. The most obvious field to point to in this regard is DeFi (Decentralized Finance), which has rapidly appropriated tens of billions of dollars of market share from Centralized Finance over the past 2 years** and shows no signs of slowing down. There is reason to believe that, as tools and infrastructure mature, this trend will bleed into other traditionally centralized verticals.

The verticals that are most suitable for disruption in this regard exhibit the following attributes:

1. They are *heavily centralized with a long tail of competitors*, such that the potential upside increases dramatically when the ability for long-tail competitors to collaborate is improved;

2. They *represent an outsized portion of economic value*, such that any increase in the total value of the industry is maximally represented in total economic activity; and

3. They are digitally native, such that transactions of economic utility are easy to capture and quantify for purposes of trustless collaboration.

At the top of this stack of viable disruptees sits Big Data. Why?

1. Today, the vast majority of data is siloed in a handful of companies which continue to build a competitive advantage, generating petabytes of data (each) each day. This, along with high salaries, attracts the most talented data scientists, who spend 80% of their time cleaning, curating, and organizing this data - they build internal tools for data management that exacerbate the divide in data quantity and quality between Big Data Monopolies and their long tail competitors;

2. The industry revenue from Big Data, officially, numbers in the hundreds of billions each year. Behind the scenes, poor data quality and bad practices cost the United States alone over $3 trillion over that same period. Data management touches nearly every company in the modern economy - if not directly, within a single degree of separation; and

3. There is not a more digitally native vertical than Big Data; value in Big Data is digitally encoded micro-assets (data points), and the tools and infrastructure used to structure, discover, contextualize them. Generating metadata for data creation is a common and straightforward task and data transformation and

processing are easily documented in a composite pipeline.

We have designed and implemented a prototype of a decentralized data marketplace on OpenSet. The marketplace uses Arweave for permanent dataset storage, and allows data consumers to reward data providers by "tipping", introducing basic economic incentives to open data marketplaces and providing an advantage over current players like Kaggle.

We will continue to build out full functionality of the marketplace, allowing data providers to monetize their data by creating token-gated datasets, easy-setup subscription licensing, and data pipeline bounties.

2 Use Cases

Building a decentralized data marketplace provides a number of advantages over traditional data marketplaces.

2.1 Data Ownership

Dataset ownership is clear and codified, and revenues generated by a dataset are distributed to the owner(s) of the dataset.

Two major mechanisms are worth highlighting in this decentralized model of data governance:

1. Data ownership tokens for individual datasets can be bought and sold. This provides quick liquidity to data providers, as well as yield farming for third party stakeholders. This also provides a pricing mechanism for data, which incentivizes data providers to generate more valuable data - datasets that are highly monetized provide higher yields to token owners, which pushes up the price of ownership tokens.

2. Fractional data ownership also lays a solid foundation for coordinating multiple stakeholders. This is important because dataset creation often requires independent parties to contribute to different parts of the data pipeline. Developing a training dataset, for example, may require aggregating data from multiple sources, multiple providers labeling that data, data transformation pre- or post-labeling, as well as the architecture and orchestration of the data pipeline.

2.2 Data Infrastructure

Building a strong fractional ownership layer for datasets provides a foundation for the first data pipelines with multiple independent parties.

Large tech monopolies have the deepest data pools, which attract the best data talent. The resulting data processing tools and infrastructure created by this talent are rarely open sourced.

We anticipate that OpenSet will unlock a significant amount of data engineering talent by providing strong economic incentives to build composable widgets for decentralized data pipelines. Creators of data pipeline widgets will charge a volume-based fee for use or accept ownership tokens of the final dataset. This is a much more meritocratic way of assembling data pipelines than currently exists in salaried roles at centralized institutions - the most experienced data engineering talent will be the most incentivized to build on OpenSet.

2.3 Data Versioning

Managing the contributions of different stakeholders in a decentralized manner during the creation and maintenance of a dataset requires that changes be recorded on chain. This means OpenSet has built in data versioning - previous versions of datasets can be reconstructed via diffs and reversible transformations, like git.

Data versioning is important for reproducibility and benchmarking. If a researcher cites a dataset and the dataset changes, it may impede the ability for other researchers to reproduce and verify results. Likewise, if version x of a computer vision model, for example, is trained on a dataset and version x+1 is trained on a modified version of that dataset, it becomes unclear whether model performance is favorably or adversely affected by the updated dataset. This is a common issue in current open data environments.

2 Conclusion

We have proposed a new economic model for a data ecosystem that leverages the transparent nature of blockchain technology to drive better pricing and collaborative practices around data generation and storage. The OpenSet ecosystem follows on a number of recent milestones in web3 - the proven success of affordable, immutable data storage; increasingly price-stable, decentralized financial mechanisms for sharing revenue among multiple independent parties; and widespread adoption by both consumers and institutions of web3 technologies.

These milestones mark a turning point in globalized commerce, beginning the transition to real transparency in global economic activity. That transition starts with transparent data practices, and a more efficient way of coordinating the economic actors involved. The foundations of enterprise activity on the blockchain are being built here. We hope that the OpenSet ecosystem will continue to thrive and function as a foundational data management layer for all enterprises and actors in our new economy.