Building on Web3:
Owning the Storage Layer

Whitepaper

Published by:
filebase

Table of Contents

Introduction

Each day, more and more of our interactions with technology involve creating or interacting with digital content and data. Being a digital content creator is now a common career option and selling user data alone is a billion-dollar industry.

Most of the time, user data is collected by different social media or eCommerce platforms without users even realizing it. Everything from user email addresses, shopping history, geographic location, or even their recent browsing history can be collected.

In 2020, it was estimated that about 64.2 zettabytes of data had been created and consumed across the Internet worldwide. According to Statista, that number is expected to exceed over 180 zettabytes by the end of 2025.

Once data is collected, it is sold to other companies to be used for things like marketing outreach emails, advertising campaigns, or personalized advertisements.

While this might be useful in some cases, most often users don't appreciate their data being sold and used for unsolicited marketing emails or ads that 'know too much'. It can appear startling to be given a personalized ad on Youtube based on your Google search history from an hour ago.

Users often ask 'how do shopping platforms know I was looking for this item?' or 'why do social media websites know everything about my history on other websites?'

The answers to these questions can be answered by the Web2 data ownership problem.

Web2's Data Ownership Problem

Web2 is often referred to as the 'read-write' version of the Internet. That's because when users interact with Web2 websites, they can read and view existing content, but also write and create their own. Content creation can include everything from a simple Facebook status or uploaded photo of your recent vacation, to content that you intentionally create for marketing your business or a community event.

In most cases, users don't think about the fact that once they upload content to social media, they don't technically own that content anymore. Technically, the social media platform now owns that content and can use any data about that content as they see fit. In some cases, they'll use that data in-house on their platform or for their own marketing purposes, but other times they'll sell that data to other third parties.

Most third-party companies that buy data are companies most haven't even heard of. Their entire business revolves around purchasing user data from Facebook, Google, Youtube, Twitter, or any other platform that collects user data that can be used for marketing and advertising. Then, they turn around and sell that data to other companies for them to use to generate new leads, personalize advertisements that result in a higher conversion rate, or curate email campaigns.

If you think about it - how else are these companies making a profit? Google is a free search engine that doesn't cost anything to use. In addition to being a search engine they do offer other paid services, like Google Drive storage, but their most popular services are free to use. By selling user data to other companies, Google can return a profit.

Another platform that is notoriously well-known for selling user data is Facebook. Facebook is entirely free for users to sign up for and use, but in 2020 alone they returned an $86 billion dollar profit. That's because everything users upload, post, or add to their profile, can be collected and sold to other companies. This includes information like users' email, date of birth, location, or phone number, but even their likes, dislikes, shopping habits, and even visual or audio content like photos and videos.

Some users don't mind that their data is sold and used in this way by other companies. What they don't realize though, is that not only is their data sold, they don't own their data. Companies can sell user data since the company owns that data - not the users.

Since users don't own their data, they have no say in how their data can be sold and used. Some platforms might allow users to limit what data is collected on them, but ultimately, the platform is still collecting and selling something on its users, even if they've opted to limit data collection on their account.

How Web2 Uses User Data

Web2 platforms and websites often use any data that it collects or purchases for a variety of workflows, including:

Marketing Campaigns

Email Campaigns

Customer Outreach

Personalized or Tailored Advertisements

Web3's Values

In comparison to Web2, Web3 is often dubbed as the ‘read-write-own’ version of the Internet. Web3 is a new way to think about and use the Internet, using technologies where data ownership is at the core of every website and application. In Web3, users own every piece of content that they generate through either posts on a Web3 platform or any data that becomes associated with their user profile. That data can’t be sold or used without the user’s permission.

This creates an environment where companies aren’t tailoring ads or marketing campaigns to each user. It also provides a market for users to sell their own data to companies, creating a new opportunity for users to monetize their digital content.

In addition to data ownership, another core value of Web3 is decentralization. Decentralization in Web3 applies to both the technologies that power Web3, such as blockchain networks and decentralized storage, but also to data ownership. The ownership of data on Web3 is decentralized since no single entity owns the data for all users like Web2 platforms do.

Web2	Web3
Centralized	Decentralized
User data is collected and sold to third parties	Users own every aspect of their own data
Central authorities control all aspects of applications, websites, and platforms	No central authority controls Web3 applications and platforms.

The Web3 Storage Layer

In Web3, the storage layer is one of the most important pieces of a fully decentralized technology stack. Data storage typically takes one of two forms across the current Web3 ecosystem:

1. Decentralized Storage Blockchain Networks

Blockchain storage networks like Sia or Filecoin store data across different blockchain nodes and provide a permanent, immutable record of what nodes have stored what data. Nodes are located across the globe in a wide variety of different geographic regions. Networks like Sia store data in several pieces across different nodes through a technology known as Erasure Coding.

Erasure coding is a forward error correction algorithm that provides data protection for distributed and decentralized storage. It is a method of data protection that breaks large amounts of information into smaller pieces, expands it with redundant data, and distributes those pieces across multiple storage locations, either drives or individual nodes. In the event that any drive or node fails or goes offline, or if data becomes corrupted, the data can be reconstructed from the segments stored in the other locations. Erasure coding helps increase data redundancy without the physical overhead or limitations associated with RAID.

Additionally, on a blockchain network, each transaction's record is publically accessible and verifiable for a transparent record of data storage.

In comparison, centralized data storage often silos data in one data center, often without any inherent replication or redundancy. This not only provides a massive risk for data loss if the data center experiences an outage or natural disaster that destroys the hardware, it also proves to be a security risk. If a hacker or bad actor gets into one portion of the data storage server, they often have access to millions of data files and entries. This can lead to data breaches where hundreds of millions of users have their private information compromised.

Platforms like Facebook, Google, Amazon, and Twitter all use centralized data storage options on the backend to store user data that they collect, along with any other information on their platforms. For other websites, applications, or users storing their own personal data, if they use centralized data storage, while it is often private, any data that they store could be collected, analyzed, and used for any number of purposes.

2. Decentralized Peer-to-Peer File Storage Protocols

In addition to data storage using decentralized blockchain networks, another alternative data storage method on Web3 involves the decentralized file storage protocol IPFS, or InterPlanetary File System. IPFS is a distributed file storage protocol. Although IPFS is used to store files, websites, data, or applications, it is not a decentralized storage network and does not use blockchain technologies on the backend.

Typically, a blockchain network uses storage resources provided by nodes on the network and nodes are incentivized to provide those resources through rewards of cryptocurrencies. IPFS is a protocol, defined as a set of rules and parameters for processing or accessing data. Protocols don’t inherently use blockchain networks or incentivize users to contribute to the protocol's success.

IPFS's Role In Web3 Data Storage

The IPFS network is a peer-to-peer storage network of nodes located across the globe, also known as IPFS peers. Each peer is capable of storing data, relaying requested data stored by other peers, or both. IPFS uses fundamental technologies such as content addressing, content linking using directed acyclic graphs (DAGs), and content discovery using distributed hash tables (DHTs).

Through these technologies, IPFS provides the ability to store files and other data that can be referred to through a content address. Content addressing refers to a form of data addressing where data is viewed or downloaded through its associated content ID, rather than its location. For comparison, the HTTP protocol uses location addressing which allows users to utilize websites with multiple pages. By using content addressing in replacement of location addressing, access to data isn't reliant on domain addresses staying online or location file paths remaining the same.

IPFS provides three important attributes to data storage in Web3:

Data Publicity: All data uploaded to IPFS is publicly available. Any file or folder's content address identifier (CID) can be accessed by anyone in the world through an IPFS gateway.

Data Ownership: Data stored on IPFS isn't owned by any central entity, since no central entity controls the IPFS network.

Data Authenticity: A file or folder's IPFS CID is generated based on the data's cryptographic hash value. Any change in the file's content, data, or metadata, will result in a new CID value. This allows for data to be authenticated and verified that it hasn't been altered or tampered with after it has been uploaded to the network.

Data Ownership in Web3

On Web3 platforms, data isn't owned by any single central entity and isn't being collected to be sold and used for marketing purposes. Web3 platforms are built with user data ownership in mind since data ownership is one of the core driving values behind Web3. By providing users with full ownership of their data, users benefit from:

Full ownership of any content they generate or create on Web3 platforms.

Full ability to monetize and distribute their content or data as they see fit.

Not being subjected to unsolicited advertisements, marketing campaigns, or emails that have been generated based on data that was sold to advertising companies.

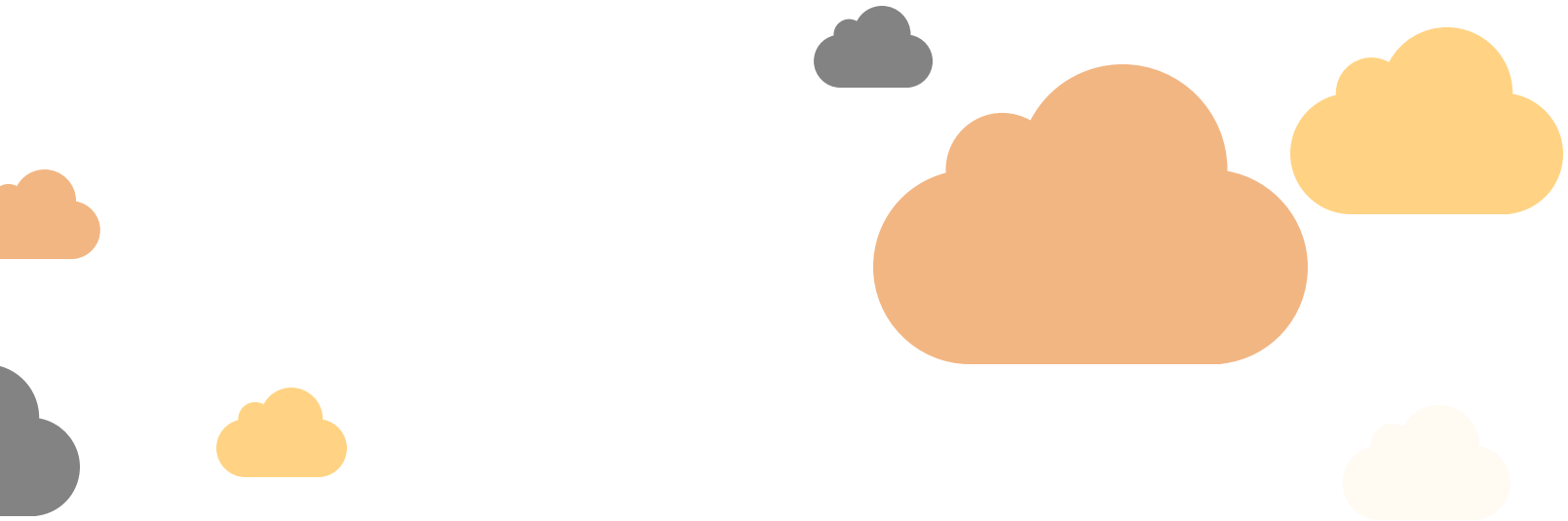
Conclusion

IPFS provides an easy-to-use decentralized data storage solution for building on Web3. IPFS allows users to store data in a manner where their data isn't collected or sold by a centralized authority. All data stored on IPFS is publicly available, providing a seamless way to share and distribute content.

There's one catch with IPFS though - unless data is 'pinned' to IPFS, it will only be stored temporarily. This can cause a problem if workflows depend on long-term data storage.

IPFS pinning is the process of storing files on IPFS nodes for long-term storage that is not cleared during the automatic garbage collection process, which is an automatic resource management process. If users host their own local IPFS nodes, pinning can be done from the IPFS desktop interface to pin files onto their local node's storage. But for users that don't host IPFS nodes themselves, they will need to use an IPFS pinning service.

IPFS pinning services pin files and folders to IPFS on behalf of the end-user and return the file or folder's CID that can be accessed and used through IPFS native URLs or IPFS gateway URLs for the end-user to view, download, and incorporate in different workflows.



Filebase: The Easy On-Ramp To IPFS

Filebase is a geo-redundant IPFS pinning service that allows you to pin files to IPFS in a secure, redundant, and performant manner across diverse geographic locations. All files uploaded to IPFS through Filebase are automatically pinned to the Filebase infrastructure with 3x replication across the globe. This ensures that your data is globally available and redundant at all times.

Filebase acts as an easy on-ramp to IPFS by offering a user-friendly web console dashboard, making drag-and-dropping files onto IPFS simple and easy. Filebase also provides an S3-compatible API for widespread integrations and configurations in current workflows. By using the IPFS Pin Sync tool, existing IPFS CIDs can be migrated to Filebase seamlessly to benefit from Filebase's 3x replication.

Next Steps

Contact Filebase Today

Learn more about how you can utilize Filebase for your specific use case.

Join our Discord Server

Chat with others utilizing IPFS and building on Web3.

Check out our Documentation

Learn more about concepts such as erasure coding or geo-redundancy, or follow Web3 tutorials for creating dApps, NFTs, or writing smart contracts.