DSA Proposal for Proof of Authorized Access

A Proposal to Limit Access at the Network Level to Authorized Parties, and Provide Immutable Logs of Data Retrieval

Discussion Paper - 6 February 2025

Introduction

This document aims to raise awareness of a security gap in the Filecoin network from a customer perspective, and inspire relevant Filecoin Improvement Proposals (FIP)[1] to address the issue.

If gaining market share from large traditional data owners is important to decentralized storage networks, the DSA proposes that key capabilities must be provided at a network level to provide sufficient security assurances for enterprises to trust this form of storage technology.

Enterprises currently consider data storage a commodity that needs to be low cost and low risk, while also meeting a list of requirements such as integration with their legacy systems, low latency, and protection against unauthorized access – all of which are difficult for the Filecoin network to provide.

That said, enterprises are also looking for new and innovative ways to monetize their data and extract the value they know lies hidden within their data stores. They are increasingly cognizant of the value of their data, with use in training large language models (LLMs) just one example. Decentralized storage systems are in a strategic position to capitalize on this opportunity – provided that certain requirements are met.

Monetization of Data

Conversations in the DSA with a number of enterprises demonstrate that certain forward-thinking organizations are seeking ways to monetize their data in ways that are efficient and trustworthy. Unlike their approach with pure data storage for internal assets, these organizations are willing to spend more and take on more risk for something that will grow their top line and create new strategic opportunities.

In order for the Filecoin network to be an integral solution for servicing this data economy, it should strive to meet the requirements specified below. Storage networks do not necessarily need to supply many of the enabling services such as NFT-based ownership credentials and/or the smart contracts that might power serviceable solutions, but it does need them to support the fundamental proofs and attestations that will both protect data as well as support data monetization.

To succeed as a web3 protocol, the Filecoin network must also allow the data owners to control their data (in the total sense that the theories around read/write/own imply) without the intervention of, or reliance on, storage providers (SPs), L2s, data preparers, and other service providers. The network itself must be able to provide sufficient attestations regarding ownership, access, and use without regards to other agents or intermediaries. This likely entails enabling token-based access control methods as well as other capabilities that properly support the control and primacy of data.

[1] https://fips.filecoin.io/

Data as an Asset Class

Data as an asset class is an important and emerging theme – not just in the context of AI LLMs but across and within the entire data universe. This theme has implications for geospatial data, integrated supply chains, pooled telemetrics, self-driving vehicles, robotics, and many other massive data collections. It represents not only a significant market opportunity but also an ample space for disrupting current technological and economic models.

The following proofs should be considered as fundamental for any decentralized storage network to support data as a tradeable asset class.

- Proof of Authenticity - This proof provides attestations that the data is immutable, that it exists in stored form in the exact same form as was originally stored. Filecoin provides this proof via the Proof of Storage (PoST) and has done so since the inception of the network
- Proof of Ownership - This proof provides attestations as to the owner of the data. It would require linking the data to some form of token-based identity that represents the data owner (e.g. with an NFT). (This proof will be outlined in a separate but related document.)
    - By way of example, an NFT could embed data CIDs so as to form an onchain immutable record that connects owners with their content
- Proof of Authorized Access - This proof supports attestations that safeguards data from unauthorized access and provides immutable data logs of authorized access. (This proof is the subject of this document.)

Additional useful capabilities include access to granular data

---

Proof of Authorized Access

A significant portion of the existing market for data storage comes from large organizations that restrict access to some data and require such access is actively controlled.

> *A core tenet is the underlying principle of data access privacy: regardless of any level of encryption, stored data cannot be retrieved by anyone on the Internet without explicit authorization, each access should be logged and capable of notification, and such access should be no less flexible than current data access approaches.*

Should the Network Provide a Proof of Authorized Access

The key question is whether the Filecoin network should provide native, built-in mechanisms to allow data owners to determine who can access and retrieve their data. This capability would be independent of intermediary actors (i.e., storage providers, and data preparers), thereby preventing malicious actions, upstream configuration errors, and other out-of-band actions that cause unauthorized exposure of data.

Today an SP cannot prevent portions of customer data from being retrieved by anyone who desires to do so. Any sense of confidentiality is provided by a concept known as ‘security through obscurity[2]’, that is, assuming others are unaware of the ease of retrieving data. If a high-profile customer discovered that their data was retrieved, the potential impact to the reputation of Filecoin could be catastrophic. In addition, details regarding access and retrievals are not a native part of the system thereby preventing auditable assurance of access integrity.

[2] https://en.wikipedia.org/wiki/Security_through_obscurity

In a desirable future state, a data owner would have the ability to control who (or what authorized system) is allowed to access any data the customer chooses to store and could modify this permission at will, not just at the time of initial data transfer to Filecoin. This control would extend to some form of auditable logs of data access thereby providing a form of ***proof of authorized access*** and not just proof of storage.

Background

The Filecoin network is currently designed as a permissionless storage network, meaning that any user can store data on it. Such data may be stored in the network across multiple data storage devices, with one or more storage devices managed by one or more storage providers (SP). Any user on the Internet today can retrieve any data stored on the Filecoin network through the use of a Content Identifier (CID) unless SPs have created their own access control mechanism.

This portion of a 2017 research report reveals that the network protocol has yet to provide security, which is a triad[3] of three principles; with data **integrity** assured through the use of CIDs, data **availability** as an option through a voluntary and optional choice of multiple replicas; but data **confidentiality** is lacking since the network lacks effective built in privacy mechanisms.

The initial Filecoin network whitepaper outlined the vision for decentralized storage. The **bold** copy in the paragraph below is currently missing from the protocol:

> *The protocol weaves these amassed resources into a self-healing storage network that anybody in the world can rely on. The network achieves robustness by replicating and dispersing content, while automatically detecting and repairing replica failures. Clients can select replication parameters to protect against different threat models.* ***The protocol's cloud storage network also provides security, as content is encrypted end-to-end at the client, while storage providers do not have access to decryption keys.*** *Filecoin works as an incentive layer on top of IPFS, which can provide storage infrastructure for any data. It is especially useful for decentralizing data, building and running distributed applications, and implementing smart contracts.*
> *– Filecoin: A Decentralized Storage Network, Protocol Labs, July 19, 2017*[4]

Network adoption is running into resistance from traditional storage users (data owners) including and especially commercial and enterprise businesses that have traditionally used on-premise storage solutions and/or, more recently, cloud-based solutions. A key aspect of these implementations is the premise a) that stored data will not be accessed (even while encrypted) by any unauthorized party and b) that data retrievals will be immutably logged for auditing and security monitoring by data owners and/or their delegates.

Filecoin is built upon the same content addressing technology that IPFS uses which means that data can be both indexed via the InterPlanetary Network Indexer (IPNI) and retrieved with IPFS (the latter providing that the Filecoin stored data does not require a retrieval price).

Data stored on the FIlecoin network may be retrieved using a variety of tools and methods – using a native Filecoin retrieval client such as Boost, Lassie, and/or go-car, as well as using gateways such as Saturn, Web3.storage, storacha.network, and, of course, IPFS. Whereas some of these systems may add their own layer of retrieval authentication and security, there is no network assurance that protects the data payloads from insecure and/or inadequate methods. Additionally, commercial-grade data owners will often store replicas with multiple SPs so as to increase data availability, thereby making securing and auditing access even more challenging.

---

[3] https://www.csoonline.com/article/568917/the-cia-triad-definition-components-and-examples.html
[4] https://research.protocol.ai/publications/filecoin-a-decentralized-storage-network/

Today the Filecoin protocol does not provide any guarantees that retrievals will only be permitted by authorized clients and assumes anyone with a valid CID can retrieve a data set easily.

The Business Case for Proof of Authorized Access

- Preserving the privacy and confidentiality of data
- Providing organizations with the ability to monitor and audit data access
- Powering new use cases around data as an asset class
- Upholding the Web3 principle that control of the data by its owner, including access.

Advantages

- Achieves goal of customer total control of their data (decentralization principle)
- Assures customers that Storage Providers are not liable to put their data at risk (mitigates the problem for customers having to conduct reviews of every SP that stores a copy of their data)
- Opens up a vast new market of potential customers that require confidentiality of their data (based on standard industry practices).
- Provides audit logs of data access (*proof of access*), not just *proof of storage*.

Disadvantages

- Modification of core network function will require time and money to design and build
- May not be technically feasible at the network protocol level

Technical Implementation

Implementation Requirements

- Must maintain the current option of public data access (full anonymous access)
- Must not impede other service providers from creating their own access control services. (In other words, openness to third party access control options but with a potential default network option.)
- Should aim for compatibility with existing SP access control services
- Must not degrade performance and scalability of the network

Open Questions

- Do we still need a better definition of ‘customer data’ through the lens of the customer, is this defined as a customer ‘file’ which can be associated with a unique CID? (Currently, customers typically think of their data as files, not CAR files, and/or they think of their data as a collection of files. “Data assets” is maybe a more appropriate term that can map to the specific distributed data storage arrangement used. This question may be outside the scope of this proposal but its resolution will help with clarity around access to customer data as manifested in the network.)
- Is it possible to build an end-to-end ‘data pipeline’ that honors access control in Filecoin and potentially into IPFS at both the initial data onboarding step (similar in concept to S3 connector) as well as management of access for the duration of stored data?

Issues Not Addressed By This Proposal

- **Access Control Methods and Interfaces –** This proposal does not recommend or prescribe any specific access control methods as these concerns are better addressed at layers above the network level. The DSA is pursuing a concurrent workstream to create a survey of access control methods being explored or implemented at the data onboarding and/or layer 2 level. Such

approaches may include AWS S3 API overlays, OAUTH overlays, UCAN[5]-based token authorization, and other access approaches.

- **Descriptions of Usage Rights** - Descriptions of usage rights such as reading, reproducing or creating derivative works from data are beyond the scope of this proposal. Such descriptions may be used to inform the design of access control systems, or as evidence in legal action against parties who make unauthorised use of stored data.

- **Digital Key Generation (DKG) and Key Management –** Digital key generation and key management is closely related to access control specifically as it relates to how private and public keys can be managed, provisioned, assigned, and recovered. This area is certainly critical to enabling widespread adoption of decentralized storage but its area of concern is at a layer further up the network stack. An example of this capability is the Lit protocol[6].

  One or more DKG and key management protocols may play a role in a token-based access control approach should that be a direction that is pursued but it will be as an enabling service to the larger and primary solution of isolating retrieval requests to only those with explicit access authorization.

- **Data Access at the Device Level** – This document's primary concern is with methods to protect the confidentiality of data and attest to access at the network level, but does not propose measures to address physical access at the device level. Whereas this is a concern with every form of data storage, other measures for device-level security (data at rest) are employed such as self-encrypting hard drives (SEDs) keyed to the BIOS or TPM of a motherboard. Other security concerns arise with data in motion (typically addressed via encrypted streams such as TLS) and data in memory (only now being explored in conjunction with growing use of Trusted Execution Environments or TEEs). All three device-level security concerns are out of scope of this document.

- **Fully Homomorphic Encryption (FHE) –** Fully Homomorphic Encryption (FHE) is a nascent technology that promises to allow for processing data without first having to decrypt it. This means companies can provide access to their data without exposing their data set. LIkewise, companies can use data provided by users and visitors without being able to see the underlying data.

  For example, users could provide attention data and purchase history in encrypted form and receive onsite recommendations without explicitly exposing this data to the site. Likewise, a data consortium such as an electric vehicle charging network, for could pool data for use by their members without having to expose the full data set to every member in the consortium.

- **Data Deletion and other Network-Level Data Functions –** This document does not recommend or suggest any direction on network-level data functions such as data deletion. While these subjects do arise in discussions with business storage users as a barrier to widespread adoption of decentralized storage, it is not within the scope of this document to explore this as a specific requirement. It should be pointed out, however, that data deletion (along with data modification) are important actions that serve as fundamental delineations in role-based access definitions.

---

[5] User Controlled Authorization Networks (UCANs) are an extension of JSON Web Tokens, designed to enable authorizing offline-first apps and distributed systems. https://www.ucan.xyz

[6] The Lit protocol is an Internet-native security and identity network. https://www.litprotocol.com/