A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions

Michel Legault

> *"Filecoin is a decentralized storage market - think of it like Airbnb for cloud storage - where anybody with extra hard drive space can sell it on the network."*
>
> Juan Benet,
> Founder/CEO of Protocol Labs and creator of IPFS

This paper provides an overview on how content can be managed with a blockchain or other distributed ledger technology (DLT), and what challenges need to be addressed in managing this content as part of transactions. Transactions on a blockchain may require supporting documents, for example, photos, reference documents, or actual contracts. As DLTs becoming an increasingly popular method to complete transactions and share information, several issues are arising that need to be addressed, such as: Where should this electronic content in documents be stored? Will the storage system have the features and functionality to properly manage this content through the "information lifecycle", including the retention and disposition of business records based on legal and regulatory requirements? The paper presents an overview of the emerging technology involved with distributed storage systems. It presents five solutions currently available, including their designs, how they secure and store files, and whether or not these files can be deleted in order to meet record disposition requirements and regulations. The discussion points out the need for alignment between multiple stakeholders and consortium members in a distributed ledger-based community with shared ecosystem scaling objectives. The challenges of scaling include the need to protect personal and sensitive information, especially when this information should normally be disposed after a record's retention period has ended.

Introduction

The technology now called "blockchain" was originally conceived in Bitcoin as a decentralized e-commerce alternative to "financial institutions serving as trusted third parties to process electronic payments" (Satoshi Nakmoto, 2008). Blockchain was meant to usher in a "trust-less" model, where mechanisms (such as cryptographic proof) could enable all parties in a distributed ledger system to reach a consensus on what the authentic data record was. In addition, the Bitcoin blockchain was meant to allow for completely non-reversible transactions.

Since 2017, Bitcoin and other alternative cryptocurrencies (known as "alt-coins") have seen tremendous growth in their popularity, now with a worldwide audience driving explosive growth in actual monetary value. Bitcoin and other cryptocurrencies have come to be seen increasingly as potential hedges against the risk of inflation and hyper-inflation (see El Salvador making Bitcoin legal tender, 2021). This is happening as fiat currencies reserves have been increased to meet economic challenges from the COVID-19 pandemic and socio-economic lockdowns.

Nevertheless, it is important to note that blockchain systems make it challenging to fully adhere to the "information lifecycle", which refers to the stages information goes through as it is managed by users, including:

- Creation/Modification
- Classification (adding metadata, identifying user access restrictions)
- Storage

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

- Retrieval/Use (through search or navigation)
- Retention and Disposition

Information, content, and data are created on blockchains. This information is also classified using the hash function, and stored directly on the blockchain of a ledger community. Information on a shared blockchain ledger is always retrievable, since users can review the details of each transaction on a public blockchain.

The paper summarizes the author's experience and professional engagement with the domain of distributed storage systems. It describes some of the challenges with storing content on blockchains, what potential solutions exist, and how distributed storage systems are part of these solutions. This paper also describes and compares a selected set of distributed storage systems currently available in the public domain, presenting the research results involving their features and approaches to immutable content.

Challenges with Storing Content on Blockchains and DLTs

The main challenge of storing content on blockchains and DLTs presents itself most transparently in the final stage of information retention and disposition. Retention of content or data is not an issue on blockchains and DLTs since this content/data is automatically immutable (that is, cannot be deleted). However, as part of the information lifecycle, a subset of content or data objects are declared as records because they are identified as containing business information that must be retained per legal and/or regulatory requirements that govern that industry or economic sector. The retention periods for records, however, usually have an end date when these records must be disposed of, unless a business record must be retained permanently based on legal and regulatory comments.

The immutability of blockchains make it challenging to destroy distributed on-chain records (Lemieux et al., 2019). Additionally, several technical capabilities that are commonly relied upon in defensible disposition plans are not yet available as part of blockchain systems (Lemieux et al., 2019). These include automated record management and classification, suspension of automated deletion, technology-assisted review (TAR), and content search of records for diligence purposes.

In addition, embedding information on blockchains has given rise to concerns about the use of blockchain recordkeeping in relation to compliance with the European Union's (EU) General Data Protection Regulations (GDPR) (Lemieux et al., 2019). Although blockchain technology enables openness and transparency in public ledgers, at the same time information recorded on-chain is permanently stored. This is the case even if a user deletes their profile, which can contain "personally identifiable information" (PII) (Hofman et al., 2019). As a result, the immutability of data stored on blockchains may conflict with EU GDPR requirements relating to the destruction of information no longer needed to meet the needs for which it was gathered in the first place. In short, blockchain technology is caught in a quandary of how to meet current data privacy rules relating to the "right to be forgotten".

Potential Solutions to these Challenges

One approach to addressing the challenge of immutability and proper records retention and disposition is to store more content off-chain than on-chain. This would allow for a more "traditional" approach where this content (including content with PII) can be stored in a document or content management solution that has functionality to enable the storage, tagging, searching, and retrieval of information, as well as the declaration, retention, and disposition of records, and the deletion of non-records (also referred to as "transitory information").

This content would be linked to related blockchain transactions through a unique URL (to this content) in a transaction's hash. There are three advantages with this approach:

- Version control: As additional versions are added, the original information block with the hash/URL will continually point to the latest version of the file, while the version history is updated and managed.
- PII data: This data, by being stored off-chain, can eventually be deleted (depending on the storage network), rather than being immutable on a blockchain where any personal information linked to the transaction could "not be forgotten".
- Records disposition: If a storage network and system allows for the deletion of files (records), then these files can be disposed of based on their retention schedules, rather than remaining as immutable data on a blockchain.

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

The Emergence of Distributed Storage Systems

In addition to traditional content management systems, people have been developing distributed storage systems, whose designs in several keys ways mirror the approach taken by blockchains and other distributed ledger systems.

A distributed or "decentralized" (this paper uses the two terms interchangeably) storage system is designed to store files across multiple file servers or locations. This type of storage system allows programs to access or store files from any network or computer. Alongside of blockchain developments, distributed storage systems are being developed by applying similar algorithms, protocols and encryption to mimic decentralized ledger technologies.

These distributed storage systems are now competing for business with more traditional server- and cloud-based storage and content management systems, such as Amazon Web Services (AWS) Content Management Systems (CMS), and Google Drive, as well as more robust content management systems, such as OpenText and Microsoft SharePoint/365. The distributed storage systems tout several advantages over traditional content management provision, such as:

• **Cost savings**: There are two aspects to cost savings:

  •Transaction fees on blockchains result from transactions when increasing amounts of data are stored on a chain or in a block. Higher transaction fees will typically inhibit a blockchain's ability to scale in a way that accommodates a large amount of community data. For example, with the Ethereum network, although it is technically possible to store data on-chain, the high fees involved make doing so impractical for most real-world use cases (Williams & Jones, 2018).

  •In terms of general cost savings, some distributed storage systems claim that because they leverage a distributed network of servers and other file storage systems, this means that they can offer storage space at a much-reduced cost compared to a set of servers controlled by centralized storage providers such as AWS.

The following table provides a cost comparison between IPFS/Filecoin and Amazon s3 infrequent access tier storage costs. It shows that IPFS/Filecoin data storage costs 38% (less than half) of the cost of Amazon's S3 — Infrequent Access per gigabyte per month. A comparison with Google cloud storage (60 TB of Google storage, which comes out to USD $0.026 per gigabyte, for example) demonstrates its costs are twice as much as Amazon's, which was already more than double that of IPFS/Filecoin (Alpha Gnome, 2021)

• Security: Distributed storage systems encrypt their data files, and store these files across the entire decentralized network, making the hacking of files and data a greater challenge compared to centralized storage and content management systems.

• Reliability: As files get distributed across a decentralized network, the risk of a single controlling "node" (more below) going down that makes files unavailable is minimized.

• Authenticity: With file storage being treated as a transaction, provenance and the authenticity of the origins of these files gets strengthened.

• Immutability: According to a review of their white papers, like blockchains and DLTs, files that are stored on these systems are immutable, meaning they are permanently stored on these systems. Although this is presented as an advantage (that is, a permanent store of knowledge that can never be lost), as previously mentioned in this paper, immutability is already a challenge for DLTs since

Table 1. Cost comparison of non-distributed and distributed storage systems

Storage Costs per GiB	IPFS/Filecoin	Amazon S3 – Infrequent Access Tier
Per Day	$0.0000018	$0.0004385
Per Month	$0.0000549	$0.0134217
Per Year	$0.0006674	$0.1603822
Average cost per deal	$0.0025003	$0.0580808

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

they cannot fulfill the disposition of records as part of their legal and regulatory requirements within the current framework.

The following section provides an overview of selected distributed storage systems with a summary of their designs and characteristics.

Examples of Distributed Storage Systems

The following distributed storage systems (platforms) were selected for this paper: InterPlanetary File System (IPFS), Arweave, SIA, Storj, and Filebase.

The InterPlanetary File System, or IPFS (https://ipfs.io/), is a peer-to-peer (P2P) distributed file system that seeks to connect all computing devices with the same system of files. While IPFS is in some ways similar to the Web, in comparative platform language, it can be viewed "as a single BitTorrent swarm, exchanging objects within one Git repository" (Benet, 2015). IPFS has content-addressed hyperlinks, encrypted content, and a data structure that allows for versioned file systems, blockchains, and even a "permanent web", which acts as a store of global knowledge. (Benet, 2015). IPFS is described by Protocol Labs itself as having "no single point of failure" and as a "trust-less" ecosystem (Benet, 2015).

IPFS is a P2P ecosystem in which no nodes are privileged (Benet, 2015). A node in IPFS means a personal computer/server that has signed up/agreed to be an IPFS storage location for content. IPFS nodes store IPFS objects in local data storage, that is, on these personal computers and servers. Nodes connect to each other and transfer objects. The objects stored and sometimes transferred include files and other data structures (Benet, 2015). The IPFS protocol is divided into a stack of sub-protocols responsible for various aspects of the system's functionality:

- Identities: manages node identity generation and verification
- Network: manages P2P connections, using various underlying network protocols
- Routing: maintains information to locate specific peers and objects
- Exchange: a novel block exchange protocol (BitSwap) that governs block distribution, modelled as a market which weakly incentivizes data replication
- Objects: encrypted content-addressed immutable objects with links
- Files: versioned file system hierarchy inspired by Github
- Naming: a self-certifying mutable name system

Although IPFS envisions a decentralized internet infrastructure upon which many different kinds of applications can be built, it currently serves the purpose being a next generation file sharing system (Benet, 2015). One notable use of IPFS was during the government's Wikipedia banning in Turkey. In this case, IPFS was used to create a Wikipedia mirror, which allowed access to Wikipedia content despite the ban (Dale, 2017).

IPFS also addresses the issue of latency, that is, delays in transmitting and/or processing data, by using the Coral distributed sloppy hash table (DSHT). Coral organizes a hierarchy of separate DSHTs into clusters depending on region and size. This enables its nodes to query peers in their local region first, thus finding nearby data without querying distant nodes (Freedman et al, 2004). This greatly reduces the latency of lookups (Benet, 2015).

IPFS recognizes that it publishes and retrieves immutable objects that are "permanent" in a digital sense. Although IPFS can track the version history of each object in the system, mutable naming of objects is not available, resulting in communication of new content happening off-band by sending IPFS links (Benet, 2015). At the same time, the IPFS console allows for users to delete files, although it is unclear if both the file and its bookkeeping information have been deleted from the storage node, or if the link has only been deleted from the console. The ambiguity involves whether or not the actual file exists in some form of limbo that, since no one else has the key to decrypt it, is essentially lost to everyone.

In its whitepapers, *Arweave* (https://www.arweave.org/) states clearly that although they have built a "monumental system of de-centralised information dissemination, we have yet to build the corresponding system of permanent knowledge storage" (Williams & Jones, 2017). Arweave thus shows its goal of creating immutable content in order to avoid failures of the past

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

where stores of knowledge have been destroyed or become un-recoverable. Arweave also refers to ongoing efforts involving censorship or manipulation of news stories by media outlets or governments after an original version is published. This might be done, for example, in order to create a "memory hole" for certain facts that may not fit a given regime's or organization's political narrative, where they are easy to conveniently forget.

Arweave acts as a browsable sister network to the internet, by providing long-term knowledge storage features that the internet needs, but currently lacks. Any web browser with the Arweave extension installed will be able to seamlessly navigate between pages stored on servers on the normal internet, and resources stored on Arweave. When pages on the normal internet are not found, the browser extension will search the "Archain" for archived copies of the page. Furthermore, Arweave is also being built to allow users to "rewind" the state of a web page and see what it looked like at a previous moment in time.

Arweave is based on a protocol where once a piece of data is stored in the data structure, it is cryptographically entangled with every other previous block in the network (Williams & Jones, 2018). This ensures that any attempt to change the contents of a document will be automatically detected and consequently rejected by the network. This allows for Arweave's claim of being able to permanently store data on-chain, "beyond the reach of accidental or intentional data loss or manipulation" (Williams & Jones, 2018).

Arweave's novel data structure, a blockweave, does not require miners to store every previous block. To achieve this, all data required to process new blocks and new transactions is "memoised" (regarding a "shadow" or slimmed-down version of the full block where the removed data can be reconstructed from other data) into the state of each individual block (Williams & Jones, 2018). Two components of a blockweave include:

- Wildfire - a system that provides for the rapid fulfilment of data requests on the network as a necessary part of participation. Wildfire works by creating a ranking system local to each node that determines how quickly new blocks and transactions are distributed to peers, based on how quickly they respond to requests and accept data from others. Peers are served by order of their rank, with poorly performing peers being blacklisted from the network entirely (Williams & Jones, 2018). This aims to address latency issues so that the Arweave solution has response speeds comparable to traditional, centralized storage providers.

- Blockshadowing - this component works by partially decoupling transactions from blocks, and only sending a minimal block "shadow" between nodes that allows peers to reconstruct a full block, instead of transmitting the full block itself (Williams & Jones, 2018).

Arweave also supports two types of archiving:

- Unverified data archiving - users can submit arbitrary information to the weave, with an associated name (an Archain Resource Locator, or ARL).

- Verified internet archiving - an internet URL is submitted to the network and a de-centralised consensus protocol is employed to agree upon its contents before storage. Verified internet archiving allows submitters to easily ensure that important information hosted on the internet will be available and accessible to them and others in the future. These backups are expected to be trustable by others in the future, as they will be guaranteed to be faithful representations of an internet URL's contents at a given time (Williams & Jones, 2018).

Arweave states that it places high value on the authenticity of the data it archives. Arweave clearly recognizes that litigation can be tied up over the authenticity of documents. In addition, in 2017, the U.S. state of Delaware signed into law amendments to Delaware's General Corporation Law to account for the use of blockchain technology in corporate recordkeeping (Lucking, 2017), which also means blockchain evidence is now admissible in court proceedings according to U.S. law (Williams & Jones, 2018). Arweave recognizes that its data archiving could speed up the verification process for authenticating records and avoiding frivolous litigation, but they do not appear to recognize the flipside of this ruling, which is that these records are immutable and can never be disposed of as part of a defensible position for records management.

The third platform, *Sia* (https://sia.tech/), has positioned itself as a "decentralized cloud storage platform that intends to compete with existing storage solutions, at both the P2P and enterprise level" (Vorick & Champine, 2014). Sia also highlights the fact that with existing centralized storage solutions, a single company

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

owns user storage data. This can in unfortunate cases lead it to "abuse privacy in the pursuit of higher profits" (Sia, 2016).

Instead of renting storage from a centralized provider, peers on Sia rent storage from each other. Sia itself stores only the "smart" storage contracts formed between parties, defining the terms of their arrangement. A blockchain is used by Sia to store these smart storage contracts. By forming a smart contract, a storage provider (also known as a host) agrees to store a client's data, and to periodically submit proof of their continued storage until the smart contract expires (Vorick & Champine, 2014).

A file that is uploaded to the Sia network is encrypted and then spread to multiple nodes across the globe. No single node contains a majority of the content of the file, but rather only small fragments. This approach, according to Sia, reduces storage costs compared to a central cloud storage provider and improves access speed and reliability (Sia, 2016).

To address potential latency issues, Sia takes a two-pronged approach:

- Clients can use regenerating codes to safeguard against hosts going offline. These codes typically operate by splitting a file into n pieces, such that the file can be recovered from any subset of m unique pieces (these values vary based on the specific code). Each piece is then encrypted and stored across many hosts, which allows a client to attain high file availability and reduced latency — for example, downloading from the closest 10 hosts, or increase download speed by downloading from the 10 fastest hosts (Vorick & Champine, 2014).

- Incentivizing hosts to maximize uptime and collect as many rewards as possible, or even larger rewards via cryptocurrency payments (Vorick & Champine, 2014).

Sia also runs into the same issue regarding immutability of files and data. In fact, Sia states that contracts do not require hosts to transfer files back to their client when requested; instead, they reward hosts for uploading files and data P2P (Vorick & Champine, 2014). Although this approach helps to bolster content in Sia's P2P network, no provisions appear to have been taken to develop a consensus that completely disposes of files and data records based on the most recent legal and regulatory requirements.

Storj Decentralized Cloud Storage (DCS) (https://www.storj.io/) describes itself as "the world's first open-source, distributed cloud storage layer that's private by design and secure by default - enabling developers to build in the best data protection and privacy into their applications as possible" (Storj, 2021). The components of Storj's framework in order to store, retrieve and maintain data include:

- Storage nodes: these distributed nodes are located across the globe, where data is reliably stored, and network bandwidth provided with appropriate responsiveness. The nodes are selected based on various technical criteria (for example, ping time, throughput, bandwidth, sufficient disk space, geographic location, uptime, history of responding accurately). A node that meets these criteria reduces latency throughout in the network and ensures high response and uptimes for users. In return for their valuable network service for the platform, nodes are paid.

- P2P communication and discovery - all peers communicate via a standard protocol where each peer provides authentication (by cryptographically proving their identity). There is complete privacy, along with the ability to look up peer network addresses by a unique identifier so that, given a peer's unique identifier, any other peer can connect to it. This creates a "trust-less" data storage and sharing network.

- Redundancy - a strategy where data is stored in a way that provides access to the data with high probability, even though any given number of individual nodes may be in an offl¬ine state. This ensures there is no single point of failure, thereby minimizing outages, downtime, bitrot, ransomware, and data breaches.

- Metadata - to track which storage nodes contain what data.

- Encryption - data is encrypted and split into 80 or more pieces, which are then stored across multiple storage nodes. If a single node goes offline, this does not block access to data, as any file sought can be reconstituted from as few as 29 of its distributed pieces that can be found in other online nodes.

- Audits and reputation - audits are used to determine

> a node's degree of stability. Failed audits result in a storage node being marked as bad, which means redistributing data to new nodes and avoiding that node altogether in the future. Such audits in turn, determine the reputation of each node. This approach also works to minimize the need for data repair.

In their whitepaper, Storj indicated that in addition to uploading, downloading, copying, and moving files, users also have the option to delete files (Storj Labs, 2018). When a user wants to delete a file, the delete operation is made, received and validated, and a signed message is returned indicating either that the storage node received the delete operation and will delete both the file and its bookkeeping information, or that it was already removed. The segment pointers to this file (regarding the metadata or key to find and open this file in the decentralized storage network) are then removed and the customer will stop being charged for that data storage.

Filebase (https://filebase.com/) is a Simple Storage Service (S3)-compatible object storage platform that allows users to "store data in a secure, redundant, and performant manner across multiple decentralized storage networks" (Filebase, 2021).

Filebase has taken a different approach compared to other distributed cloud storage (DCS) networks. Filebase allows users to select a DCS - either the Sia, Storj, or Skynet DCS —as their storage layer. Filebase leverages unused storage capacity and rents storage from these DCS networks, managing all smart storage contracts on behalf of users, which serves as a cryptographic Service Level Agreement (SLA).

The Filebase platform includes mechanisms for high-availability, redundancy, and privacy. When servers on these networks go offline, data is automatically repaired and uploaded to new hosts, providing for minimal latency and interruptions. Filebase claims it can achieve 3x redundancy for every object (Filebase, 2021).

Unlike other leading DCS networks, Filebase has no requirement to generate or purchase cryptocurrency as part of its service since it has no token. Filebase appears to be positioning itself as an intermediary between the DCS networks for users seeking distributed storage for their files.

From a retention and disposition perspective, Filebase allows users to delete uploaded files in their folders ("buckets"). Based on a review of documentation available on their website, Filebase does not appear to clearly explain anywhere if requests are sent to the selected DCS network to permanently delete a file, along with its "bookkeeping information". At the same time, if the only link to this file in the DCS is the link provided on the Filebase interface and console, this file may be forever "lost" without the ability to decrypt it or to identify its owner. This scenario for Filebase as a DCS network intermediary needs to be better understood to see if a defensible position for records disposition can be established.

Comparison of Distributed Storage Systems across the Information Lifecycle Stages

The following table provides a comparison of the distributed storage systems reviewed above, including how they relate to the "information lifecycle" stages, as well as specific attributes within each of these stages.

After having made this comparison, I make no recommendation in this paper for any one of these distributed storage solutions as optimal for content storage off-chain. Each has their own strengths that favour different uses:

- IPFS is one of the original protocols and has the size and features to be leveraged by larger organizations. Sia, Storj, and Filebase are also vying for market share with organizations (from small to large) and not just individual users, but they are relative newcomers compared to IPFS.

- Arweave has positioned itself as a permanent store of knowledge, and organizations should consider this solution particularly for data that requires permanent archiving of content.

Users and organizations must clearly define and document their content management requirements and compare these to the features of each solution in order to select the right solution for their unique needs.

Conclusion

With the rise of blockchain and DLTs, an increasing need has arisen to understand what data should be stored on-chain and what data is best stored off-chain. Data that contains personal identifiable information (PII) and/or needs to be disposed of after a defined retention period should not be stored on-chain whenever possible. This is because that data will then

Table 2. A comparison of distributed storage systems

Information Lifecycle Stage	IPFS	Arweave	Sia	Storj	Filebase
Creation/ Modification (including version control)	No barriers to content creation; allows for version control when updating files	No barriers to content creation; separate files need to be added for manual tracking of versions to maintain the immutability of each version	No barriers to content creation; no indication that version control exists (re: allowing for previous versions to exist as part of a version history for the file)	No barriers to content creation; no clear indication on the ability to manage versions of files	No barriers to content creation; no clear indication on the ability to manage versions of files
Classification (including metadata)	Metadata captured for pointers; no other indication of being able to add metadata beyond the file name	No other indication of being able to add metadata beyond the file name	No other indication of being able to add metadata beyond the file name	Metadata captured for pointers to the file in the network; no other indication of being able to add metadata beyond the file name	User-defined metadata can be added to a file
Storage (including security and authenticity)	Protocol ensures encryption of files for secure storage and unique identifier to confirm authenticity	Protocol ensures encryption of files for secure storage and unique identifier to confirm authenticity	Protocol ensures encryption of files for secure storage and unique identifier to confirm authenticity	Protocol ensures encryption of files for secure storage and unique identifier to confirm authenticity	Protocol ensures encryption of files for secure storage and unique identifier to confirm authenticity; also allows for content to be either Private or Public
Retrieval/Use	Unique URL for a file can be shared with other users	Users can share and link to Arweave resources like	Files can be downloaded; files can be shared publicly and privately; user needs	Files can be downloaded; user can share a bearer credential with other	In addition to the ability to set content as either Private or Public, a

Table 2. A comparison of distributed storage systems (cont'd)

Information Lifecycle Stage	IPFS	Arweave	Sia	Storj	Filebase
		normal web addresses (provided Arweave is enabled on the user's web browser)	to share the "siafile" (which includes metadata pointers for the file) with others so they can download the file	users to provide them access to a file	unique URL for a file can be shared with other users
Retention/ Disposition	Console allows for the deletion of files; unclear if file is also deleted from the nodes or just from the console	Files are permanently stored	Console allows for the deletion of files; unclear if file is also deleted from the nodes or just from the console	Ability to delete a file and its bookkeeping information	Console allows for the deletion of files; unclear if file is also deleted from the nodes or just from the console

become immutable, which in turn makes it more difficult for someone to have the ability to "be forgotten" in cyberspace.

For users and organizations that want to extend the paradigm of "decentralization" to file storage, the development of distributed storage systems offers an interesting alternative to traditional, more centralized on-premise and cloud storage providers. Distributed storage systems are based on blockchain protocols. These systems offer interesting alternatives to more traditional content management and storage systems, as they offer more secure storage and authentication through encryption and pointer metadata, respectively. They also promise reduced costs by leveraging a network of distributed nodes, such that no new hardware or server costs are needed for these systems to provide storage.

At the same time, these new distributed storage systems have their own challenges. If one of these systems also creates immutable copies of files, it presents a challenge to protect PII and dispose of records. In addition, with these being relatively new systems, other aspects of information management are still maturing at the same time, such as user-provided metadata on files, version control, and seamless user access for multiple users.

File storage will continue to be a topic of interest in the blockchain and DLT space. In particular, the recent growth of non-fungible tokens (NFTs), which are now associated with content such as books, music, and artwork, attests to the need for secure storage of these tokens. Distributed storage systems have an important role to play in developing decentralized ecosystems. Their increasing technological maturity is likely to continue to disrupt the file storage and content management industry.

Additional discussions need to be held to better vet the requirements that organizations have for the storage, retention and disposition of their content, and how these distributed storage solutions either meet or do not meet these requirements. Meeting both the institutional requirements as well as the social preconditioning for onboarding new technologies will be key for distributed storage solutions to become serious rivals to existing and well-established content management systems.

A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions *Michel Legault*

References

Benet, Juan. 2015. *IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3)*. DOI: https://ipfs.io/ipfs/QmR7GSQM93Cx5eAg6a6yRzNde1FQv7uL6X1o4k7zrJa3LX/ipfs.draft3.pdf

Dale, Brady. 2017. Turkey Can't Block This Copy of Wikipedia. *Observer Media*. DOI: https://observer.com/2017/05/turkey-wikipedia-ipfs/

Filebase. DU. *Introduction*. DOI: https://docs.filebase.com/

Freedman, M.J., Freudenthal, E., & Mazieres, D. 2004. Democratizing Content Publication with Coral. *NSDI*, Volume 4.

Hofman, Darra et al. 2019. The Margin between the Edge of the World and Infinite Possibility: Blockchain, GDPR and Information Governance. *Records Management Journal*, 29, no. ½. DOI: https://doi.org/10.1108/RMJ-12-2018-0045

Lemieux, V., Hoffman, D., Batista, D., Joo, A. 2019. Blockchain Technology & Recordkeeping. *ARMA International Educational Foundation*. DOI: armaedfoundation.org/wp-content/uploads/2019/06/AIEF-Research-Paper-Blockchain-Technology-Recordkeeping.pdf

Lucking, D. 2017. *Delaware Passes Law Permitting Companies to Use Blockchain Technology to Issue and Track Shares*. DOI: https://www.allenovery.com/en-gb/global/news-and-insights/publications/delaware-passes-law-permitting-companies-to-use-blockchain-technology-to-issue-and-track-shares

Nakamoto, S. 2008. *Bitcoin: A Peer-to-Peer Electronic Cash System*. DOI: https://bitcoin.org/bitcoin.pdf

Sia. 2016. *Sia Introduction*. DOI: https://youtu.be/B4YGpWxyn6Y

Storj Labs. 2018. *Storj: A Decentralized Cloud Storage Network Framework*. DOI: https://www.storj.io/storj.pdf

Storj. 2021. *Introduction*. DOI: https://tardigrade.gitbooks.storj.io/

Vorick, D., & Champine, L. 2014. *Sia: Simple Decentralized Storage*. DOI: https://sia.tech/sia.pdf

Williams, S., & Jones, W. 2017. *Archain: An Open, Irrevocable, Unforgeable and Uncensorable Archive for the Internet*. DOI: https://www.arweave.org/whitepaper.pdf

Williams, S., & Jones, W., 2018. *Arweave Lightpaper Version 0.9*. DOI: https://www.arweave.org/files/arweave-lightpaper.pdf

About the Author

Michel Legault has over 17 years of experience in Information Technology and Information Management, with particular expertise in Information, Knowledge, Content and Records Management. Michel's particular expertise is with strategy, governance, processes and solutions, and project management. Michel has additional expertise with Information Architecture. Michel is a certified Project Management Professional (PMP), an OpenText Content Server Business Consultant, and an AIIM Enterprise Records Management (ERM) Specialist. Michel has also completed an introductory certificate in blockchain / cryptocurrencies from the University of Nicosia. Michel has a wide range of experience in different industries, including the Public and Non-Profit Sectors, Transportation, Energy and Resources, the Life Sciences, Financial Services and Consumer Products. Michel was a co-author for the Deloitte paper 'The digital workplace: Think, share, do - Transform your employee experience" (2011), and has made presentations on the following: "Information Governance in The Age of Blockchain" (ARMA NCR Conference, November 2018), "Ying and Yang: Governance for Structured and Unstructured Content" (ARMA Canada Conference, May 2017), and "Functional Classification and Records Management in the Ontario Public Service" (IMAPS Symposium, May 2012).

Citation: Legault, M. 2021. A Practitioner's View on Distributed Storage Systems: Overview, Challenges and Potential Solutions. Technology Innovation Management Review, 11(6): 32-41. http://doi.org/10.22215/timreview/1448

Keywords: Distributed storage, blockchain, decentralization, information lifecycle