



The Ethereum Wayback Machine

A Solution to the Long-Term Blockchain Transaction Data Availability Problem.

Abstract

We propose the Ethereum Wayback Machine (EWM), a protocol for verifiable access to blockchain data. The EWM is a set of open-source specifications, source code and data artifacts for programmatic access to historical, archival transaction data without centralized rent-seeking intermediaries. As Ethereum moves towards a rollup-centric roadmap[1], there has been a proliferation of rollup technologies and applications atop, enabled by the increased execution and throughput limits of rollups. The new limits in effect lead to an increase in the state growth, making way for technologies like data availability protocols and others to combat the state growth. In this new world, access to historical data becomes harder, less secure and more centralized. The EWM as a component of the Covalent Network is the indispensable ally in ensuring long-term data availability of blockchain transaction data.

1. From Full-Nodes to Rollups

In the early days of Ethereum, from inception of the Ethereum Mainnet in 2015 to 2019, a single node was sufficient to handle all onchain operations. The Ethereum network was relatively small and the demands of applications on the “World Computer” were manageable with the available throughput (i.e., transactions per second). This was the “Full-Node Era” and marked the beginning of Ethereum’s journey, as it laid the foundation for decentralized applications and smart contracts. As there were only a few applications live on Ethereum at the time, the historical state was small and a single-node architecture was enough to verify transactions and propagate the network forward. Moreover, you could run a node with the full historical archive of the Ethereum network (i.e., a full archive node) on modest hardware and use it to query historical data for record-keeping.

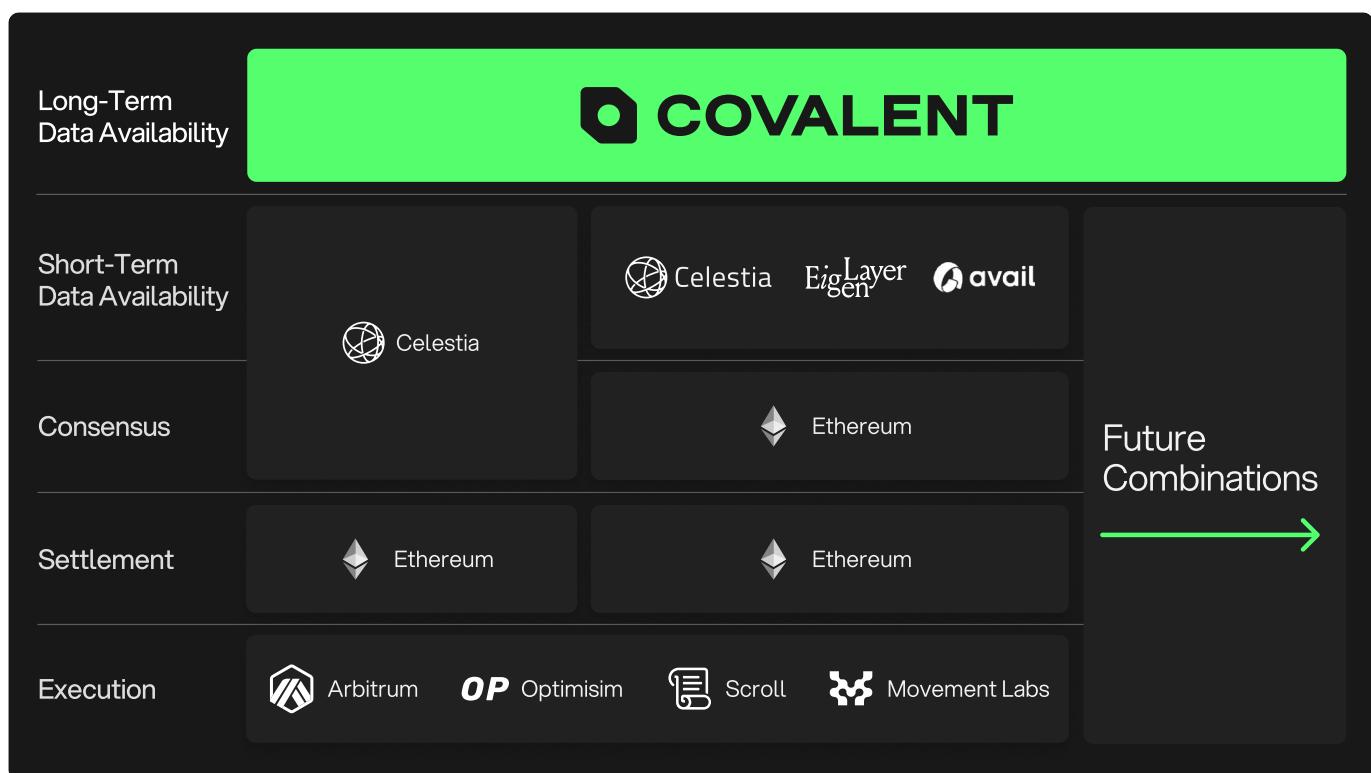
1.1 The Light-Client Era

As Ethereum gained popularity and adoption surged, the demand for scalability became evident. Ethereum entered the “Light-Client Era”, spanning from 2017 to 2022. As the historical state of the Ethereum network started to grow, it was getting increasingly difficult for the average user to run a full archive node with all of the historical data. The resource demands of running a full node ran counter to the ideals of decentralization and the ethos espoused by the Ethereum community. For example, the most recent client software versions for Ethereum require SSDs to fully synchronize the Ethereum blockchain: a regular hard drive will not suffice and cannot keep up with the required IO. Only businesses like validator pools, exchanges, indexers, and block explorers run full nodes today. At a much smaller scale, individuals run a full node for altruistic reasons to help propagate the network. To make running nodes more accessible, light client protocols were introduced to reduce resource requirements. A light client, unlike a full node, does not directly interact with the blockchain, they instead use full nodes as intermediaries.

Today every single client software for Ethereum (e.g.: Geth) supports a light client protocol. As light clients became a popular way to participate in the Ethereum network, full nodes became few and far between. Specialized providers like RPC vendors (Infura, Alchemy, Chainstack, etc.) and indexers like Covalent are the only places to get historical data today. As Ethereum gained popularity and adoption surged, the demand for more throughput was evident. A light client only helps with network decentralization, not with execution throughput. Enter the “Rollup Era”.

1.2 The Rollup Era

Rollups have become the centerpiece of Ethereum’s scaling roadmap[1], by executing transactions and storing data off-chain. Rollups, such as Optimism and Arbitrum, batch execute transactions off-chain and periodically commit to the Ethereum mainnet. By moving execution and storage off-chain, the Ethereum network can scale throughput and reduce congestion. However, as with every class of scaling solutions, either on the decentralization front or on execution throughput, a new problem arises because of rollups: data availability.



2. What is Data Availability?

Data availability, in this context, revolves around making transaction data accessible for validation by other participants in the network. This process is as follows:

1. Transactions are requested and entered into the mempool.
2. The proposer orders these into a new block.
3. The new block is broadcasted to the network, aiming for a 60-66% acceptance rate on the proposal before widespread propagation. However, all nodes still re-execute it for verification, although this is not needed for consensus.
4. Validators download the new block and re-execute each transaction in the block. They can accept or reject the block by comparing the proof embedded in the block and their executions. By doing so, they can check that the block produces the expected results.
5. The block achieves consensus, and the network begins building more blocks on top of the previous block. It's important to note that consensus is typically reached quickly, and the network moves forward with minimal delay.

In the context of blockchain networks, ensuring data availability is a paramount concern, focusing on short-term and long-term access to transaction data. In practice, the term 'withholding attack,' which refers to the malicious act of not making transaction data available for validation, has been a topic of discussion. However, it is worth noting that while short-term data availability is crucial for the immediate health and security of a network, it's not the only concern. To ensure that the state growth is manageable, most data availability protocols only keep the data around for about 2 weeks.

The rise of Layer 2 solutions has brought about a growing need for high liveness and retrievability guarantees for consensus regarding the calldata. This need is especially acute in cases where validity proofs are required, as the integrity of these proofs heavily relies on data availability. We term this as short-term data availability critical to the health and security of propagating a network. Many novel protocols like EigenLayer[2], Celestia[3] and Avail[4] are designing data availability networks to address the needs for consensus and validation. These evolving data availability protocols are designed to ensure that data is available when required, reduce the risk of withholding attacks, and support the needs of various applications, including those relying on L2 validity proofs.

To ensure that the state growth is manageable, most data availability protocols only keep the data around for about 2 weeks.

In contrast to the short-term data availability requirements, we coin the term long-term data availability. Long-term data availability is the ability to access the transaction data long after consensus and finalization. This aspect of data availability is arguably more important for applications building atop the network and is a responsibility quite distinct from running and propagating a network.

3. Ethereum's Roadmap Makes Long-Term Data Availability Inevitable

We define long-term data availability as the sustained accessibility and availability of all historical blockchain data over extended periods. While Ethereum's historical data has always been available directly through the client, this is changing. Due to Ethereum's adoption of rollups and other protocol improvements like Proto-Danksharding (EIP-4844)[5], Danksharding (where data blobs required for validation are deleted after 1 to 3 months) and state expiry (EIP-4444)[6], the responsibility of providing historical data moves outside the Ethereum core protocol. As the network evolves, ensuring continued access to historical blockchain data over extended periods becomes imperative for various applications such as taxation, auditing, AI models, and regulatory compliance. Balancing the need for scalability with the preservation of historical data presents a significant challenge that the Ethereum community must address to maintain the network's utility and usability.

3.1 Why Should One Care About Long-term Data Availability?

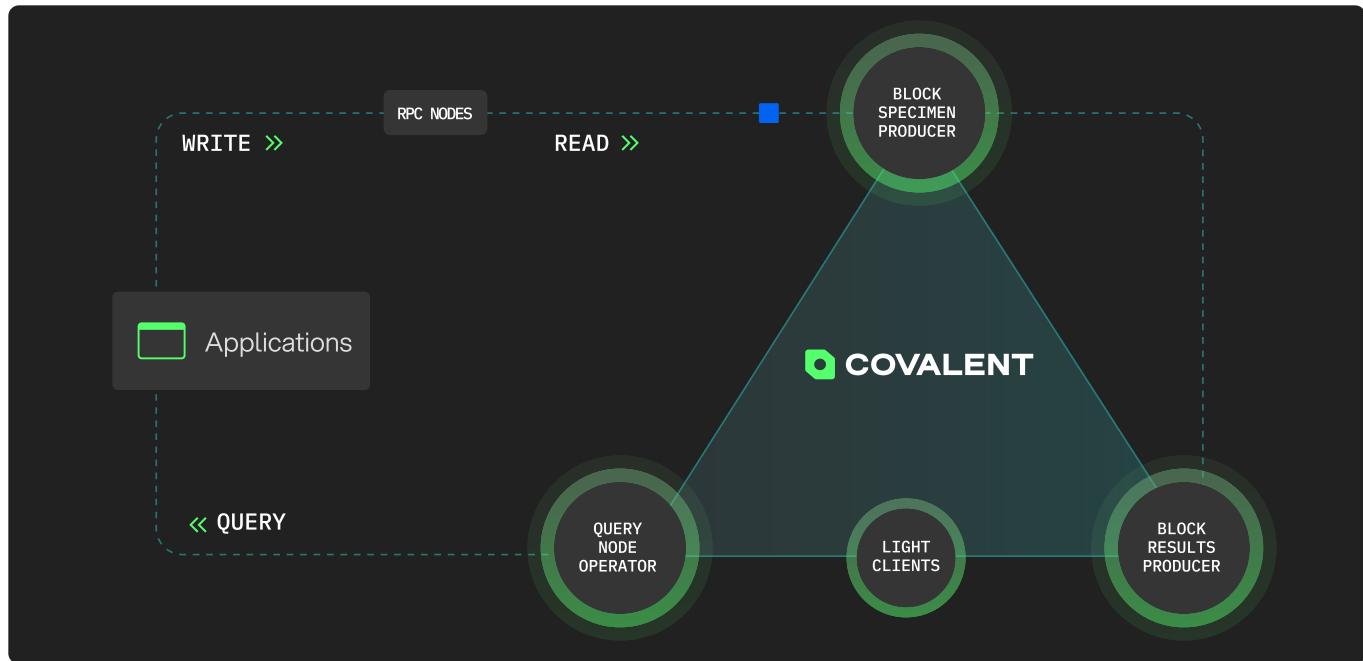
Long-term data availability is critical for several reasons, namely:

- 1. Transparency:** It allows for the full historical transactions record to be accessed and analyzed by anyone who wants to read blockchain data.
- 2. Regulatory compliance:** Regulators may require access to historical data to investigate potential fraud or other illegal activities.
- 3. Investment strategy:** Investors and stakeholders may require access to full historical data to make informed decisions about investments or other blockchain-based activities.
- 4. Adoption:** The lack of data availability can hinder the adoption of blockchain technology as developers may be hesitant to build on a network where historical data is not fully accessible or may require significant effort to access.

All of the above can lead to a reduction in the number of applications and use cases for blockchain technology, limiting its potential impact and adoption. This is where the Ethereum Wayback Machine (EWM) comes in as Ethereum's answer to its long-term data availability issue.

4. The EWM Plugs Into the Covalent Network

Recognizing the challenge of long-term data availability in Ethereum, we've taken the step of open-sourcing the specifications, source codes, and the reference implementation that powers EWM. EWM stands out as a noteworthy departure from traditional archive nodes. Its uniqueness lies in the extent of its capabilities, guaranteeing that historical Ethereum data stays accessible, verifiable, enriched, indexed, and queryable. It's not just a passive data storage solution; it actively ensures the accessibility and usefulness of the data.



1. Block Specimens Producer (BSP): a novel invention to securely extract data from blockchains with cryptographic proofs. The BSP exists as patches atop the popular Ethereum clients like Geth. The BSP can run in historical extraction mode for archival data, or synchronous mode for live data right after block execution within a client context, outputting to a storage format known as Block Specimens.

2. Block Results Producer (BRP): a concurrent block re-execution framework for enriching Block Specimens. One of the many outputs is a Block Result, a one-to-one representation of block data returned from an RPC call to a blockchain node with optional and additional fields.

3. EWM Light Client: A lightweight, decentralized client designed to validate the existence and integrity of Block Specimens within the Covalent Network.

As Ethereum embraces a modular design, the Covalent Network takes on the responsibility of ensuring access to long-term blockchain data for indexing and querying.

By offering a decentralized network, developers and other ecosystem participants can use the BSPs and the BRPs to rebuild a full canonical representation of a blockchain, host a database

with normalized schemas, or seamlessly participate as a query node for the GoldRush API. There are several features that make the Covalent Network stand out as it makes long-term blockchain data available for indexing and querying:

4.1 Decentralized

There is concern within the Ethereum community over censorship risks with centralized organizations providing historical data. We agree with this risk and it is one of the primary reasons why we are decoupling our solution to form a decentralized network.

4.2 Removing the RPC Bottleneck

Thanks to two novel technologies on the Covalent Network, the Block Specimen and Block Results, the network will not be burdened by the limitations that come with querying the RPC layer for data. Any protocol building on top of the Covalent Network will be able to scale freely without the constraint of running nodes.

4.3 Secured via Proofs

To ensure the integrity of the data made available over the Covalent Network, cryptographic proofs are employed. For every piece of work completed, such as creating a Block Specimen or responding to an API query for a developer, a respective proof is created.

The Covalent Network relies on the integrity of these proofs as they serve as evidence of the data's correctness. The architecture of the Covalent Network mandates that these proofs match the majority of hashes for the same block. In practice, this means that the data is cross-verified by multiple parties, and the system ensures a consensus on the accuracy of the data before making it available. This consensus driven approach minimizes the risk of erroneous or malicious data entering the network, thereby preserving the integrity and trustworthiness of historical blockchain data.

4.4 Enrichment

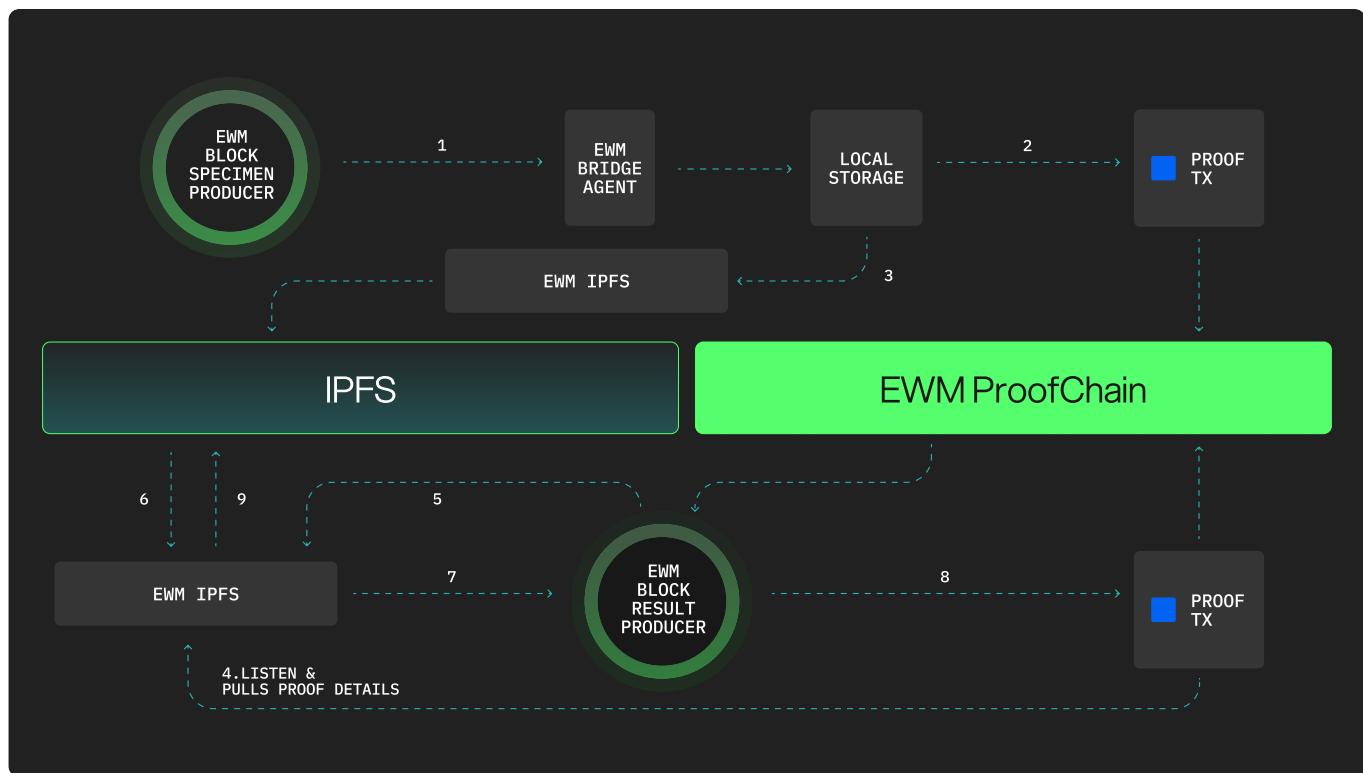
Furthermore, the Covalent Network offers enrichment capabilities, like contract tracing, shadow event annotations, and off-chain NFT metadata, empowering developers and users with extensive insights and analytics derived from comprehensive historical data.

4.5 Modular and Composable

Covalent Network operators can build canonical representations of any blockchain, as the BSP technology is simply a patch designed for client software.

Therefore, any validator can continue validating Ethereum while providing live or historical data to the Covalent Network. Furthermore, if a non-Ethereum network running an EVM execution environment wishes to adopt the BSP technology, they can fork the patch for their client and begin exporting data to the Covalent Network. It simply complements any stack.

5. The EWM in Action



Components of the Covalent Network, namely the BSP, have been operational in production for over a year. The BRP went live in the third quarter of 2023. BSPs currently submit proof transactions to a Proof Chain contract on the Moonbeam blockchain, though any EVM-compatible blockchain will work. Each transaction emits a

Components of the Covalent Network, namely the BSP, have been operational in production for over a year. The BRP went live in the third quarter of 2023. BSPs currently submit proof transactions to a Proof Chain contract on the Moonbeam blockchain, though any EVM-compatible blockchain will work. Each transaction emits a

[BlockSpecimenProductionProofSubmitted](#) log event with the following fields: [chainId](#), [blockHeight](#), [blockHash](#), [specimenHash](#), [storageURL](#), and [submittedStake](#).

Similarly, there is a [BlockResultProductionProofSubmitted](#) log event for each Block Result after re-execution. Any interested party can crawl the blockchain for these events and download the corresponding file pointed by [storageURL](#) to construct the canonical representation of the respective blockchain.

6. CXT: The Currency of the Covalent Network

The Covalent Network's responsibility as the EWM has far-reaching implications for not just the ecosystem but the underlying infrastructure currency of the network: CXT. CXT, an ERC20 token on Ethereum is the Staking and Governance token for the Covalent Network.

As developers increasingly rely on the platform for historical Ethereum and rollup data, CXT becomes indispensable in accessing this information. The token facilitates API queries, driving its functional usage and creating organic demand. Meanwhile, the staking mechanism ensures data integrity with network operators and other participants. This incentivizes responsible behaviour and ensures the data provided over the network is verifiable.

More importantly, the Covalent Network introduces an incentive structure that is unique and essential in making sure it can serve as the EWM by ensuring the platform attracts the most appropriate and skilled data providers. By becoming a Network Operator, one can export raw blockchain data with additional fields to their own data lakes or warehouses. They can then serve this data on their platform or utilize it for in-house analysis all while earning CXT so long as they fulfill Network Operator duties on the Covalent Network. This structure fosters data integrity, supports robust data-sharing practices, and ensures a reliable and accessible repository of historical blockchain data.

7. Looking Ahead to the Vibrant Data Future

In the emerging web3 landscape, the integration of Ethereum rollups for scalability and the Covalent Network for efficient data retrieval promises to revolutionize the ecosystem. With Ethereum rollups enabling a significant increase in transaction throughput and the Covalent Network serving as a robust data querying solution, true scalability is achieved both for writing and reading data on the Ethereum blockchain. This transformative combination paves the way for better infrastructure to build decentralized applications (dApps) and, in turn, leads to the development of more powerful and user-friendly apps that can attract a broader audience. Ultimately, the synergy between the EVM and EWM drives mass adoption, as developers can now harness the potential of Ethereum without facing the bottlenecks that were hindering its scalability, making it an ideal platform for building innovative and sustainable web3 solutions.

8. Conclusions

1. As Ethereum enters the “Rollup Era” on its path toward improved scalability, the Covalent Network stands as an indispensable ally in ensuring long-term data availability.
2. With its decentralized approach, ability to overcome the RPC bottleneck, modularity, and verifiability through proofs, the Covalent Network plays a pivotal role in enabling developers access to full historical data for Ethereum and other base layers.
3. In the ever-evolving landscape of blockchain technology, the EWM as a component of the Covalent Network promises a brighter future, where historical data remains accessible and blockchain applications thrive.

References

1. A rollup-centric Ethereum roadmap
<https://ethereum-magicians.org/t/a-rollup-centric-ethereum-roadmap/4698>
2. EigenLayer: <https://www.eigenlayer.xyz/>
3. Celestia: <https://celestia.org/>
4. Avail: <https://www.availproject.org/>
5. Shard Blob Transactions: <https://eips.ethereum.org/EIPS/eip-4844/>
6. Bound Historical Data in Execution Clients: <https://eips.ethereum.org/EIPS/eip-4444/>

About Covalent

Covalent is the leading modular data infrastructure layer dedicated to solving major challenges in blockchain and AI, including verifiability, decentralized AI inference, and Long-Term Data Availability. Its large reservoir of structured, verifiable data enhances decentralized training and inference, reducing the risk of manipulated or biased AI models. Additionally, the Covalent Network's Ethereum Wayback Machine ensures secure, decentralized access to Ethereum's transaction data. Trusted by over 3,000 leading organizations, Covalent powers AI, DeFi, GameFi, and more with unfettered access to on-chain data from over 200 blockchains.



Modular Data Infrastructure for AI.

X [@Covalent_HQ](#)

Telegram [covalenthq.com/telegram](#)

Discord [covalenthq.com/discord](#)

Website [covalenthq.com](#)