IBM Storage

Why new off-chain storage is required for blockchains

Document version 4.1

Contents

List of figures

INTRODUCTION

In pursuing the storage point of view for blockchain technology, the test team has come up against a recurrent theme. This recurrent theme is that no new storage is needed for off-chain data as most companies are already using the data to be utilized by the blockchain. Because the companies are already using the data, it already exists somewhere in their storage environment and only needs to be referenced by the blockchain application programming interface (API). This paper examines this to see if it is valid.

1 What a blockchain is, and what it is not

A blockchain is a distributed ledger that is utilized by a consortium of related users. These users might be a group of banks, a group of lending institutions, or a polyglot of food providers. Each will have a full copy of the distributed ledger with its links to any off-chain data. The off-chain data must be sharable and accessible for the concept of a distributed ledger to work.

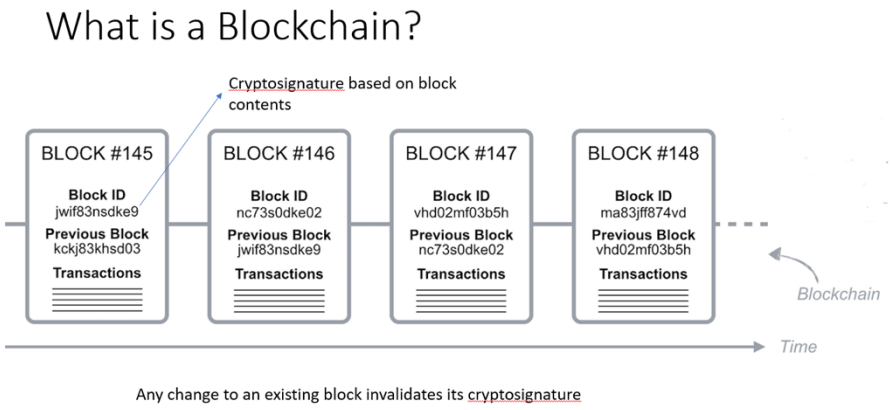


Figure 1: Example blockchain structure

1.1 Costs of blockchain technologies

The same calculation to deliver the amount of storage can also be applied to the cost of various blockchain technologies. Using published cost per transaction data for bitcoin and Ethereum, you can compare this to a five-node IBM Hyperledger system.

1.1.1 Hyperledger costs

The current base enterprise-level cost for an IBM Hyperledger is USD 1000 per month plus an additional USD 1000 per active node bringing the cost to USD 6000 per month. This does not include other software as a service (SaaS) costs that the project may incur as those will vary between projects.

1.1.2 Non-permissioned blockchain costs

At the time of this writing, the cost per transaction for bitcoin is USD 1.30 per transaction. However, this might change based on the cost of the underlying bitcoin cryptocurrency. The cost per Ethereum transaction is USD 0.25 per transaction, and this again depends on the cost of the underlying Ether cryptocurrency.

1.1.3 Results

The costs for non-permissioned blockchain vary by transaction rate and the current value of the underlying cryptocurrency. The cost of the permissioned blockchain such as Hyperledger vary by number of nodes and the amount of SaaS used in the application. This results in a fixed cost per month for permissioned blockchain, but costs per month for non-permissioned could vary widely. Based on these charges per transaction, the cost for using the various blockchain platforms at the current cost per transaction is easily calculated.

Figure 2: Cost per month for blockchains

As shown in Figure 2, the cost of the various non-permissioned blockchains such as bitcoin and Ethereum get excessive beyond, or even at, a low transaction per second rates.

1.2 What then is Offchain Data?

Off-chain data is any non-transactional data that is too large to be stored in the blockchain efficiently, or, requires the ability to be changed or deleted. Figure 3 shows examples of offchain data types.

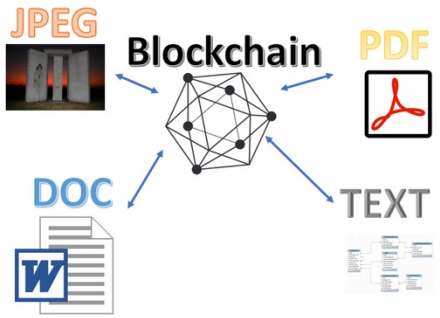


Figure 3: Example off-chain data

1.3 All of this data must be shared

For example, a blockchain is not a silo application where business A sees only business A's data. This is where the belief that data reuse from existing systems seems to have root. To be on the blockchain, or referenced by the blockchain, the data must be shared among other members of the consortium. Figure 4 is a graphic representation of an extended data set of offchain data on various peer-nodes.

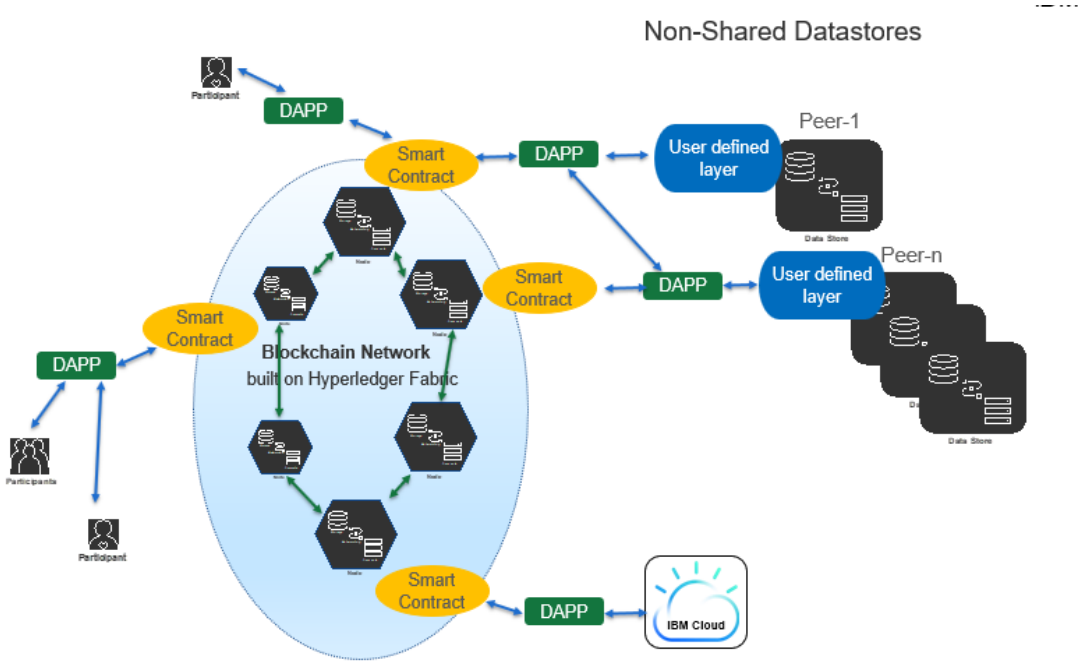


Figure 4: Example blockchain data flows

1.4 On-chain storage calculations

Given a basic understanding of storage and some basic understanding of how blockchain stores data, you can derive a back-of-the-envelope calculation on how much storage is required for Hyperledger blockchain.

1.4.1 On-chain blockchain assumptions

The following assumptions are made in the on-chain storage calculations:

1. All calculations will be in actual bytes, that is 1024 is 1 KB.
2. All hyperledger blocks are 1 megabyte (MB)
3. Only hash, signature, or key data is stored in the blockchain
4. The company works 8 hour per day
5. The company has roughly 240 processing days per year and the number of transactions per second is averaged over those days only (allows for holidays and weekends.)
6. Currently bitcoin is storing about 1400 basic transactions per MB block. Hyper ledger has larger headers and a more robust transaction size so 1000 transactions per block was selected. Blockchain transactions currently run about 5 KB each which equates to 205 TPS.

1.4.2 Calculation

Using the above assumptions, the calculation to find out the amount of storage required per TPS becomes:

(1 TPS/1000 TB)*1024 KB*3500 sec/hr*8 hr/day*240 days/year=
7,077,888 KB of data per transaction per year

6,912 MB = 6.75 GB = .00659 TB/transaction/yr

Because the transaction sizes might change according to the type of transaction being stored in the blockchain, the chart has been prepared using various transaction sizes.

2 The issues with using existing data stores

There are several key issues with using existing data stores with new blockchain-based projects.

2.1 Access issues

What would happen if instead of providing an isolated data store that holds the off-chain reference data, each of the consortium members had to provide links to their existing data stores? Each consortium member would have to finance the programming of the various data connectors, encryption, and security applications needed. While it might be perfectly okay for fellow consortium members to see document XYZ, it is not for them to possibly see document ZZX which is not on the blockchain.

2.2 Security issues

When using existing data stores, there is a high probability of accidental disclosure of confidential data. You must also consider that the very shared nature of the blockchain environment might lead to deliberate hacking into the consortium members individual data stores by a malicious actor facilitated by the blockchain access paths.

2.3 Cryptographic issues

Even if an existing data store could be used, it needs to be modified to store the required hash, cryptographic codes, and keys needed to provide proof to the blockchain that the data has not been modified since last access. For already active data stores that are modified frequently, this recalculation and storing of hash and cryptographic keys could become onerous.

2.4 Performance issues

Another issue with using existing data stores is the problem with getting uniform response times from multiple, independent, data stores using different database technologies, storage hardware, and server capabilities across the consortium. A unification of hardware, software, and data store technology is required at the consortium level to meet the needed service levels.

2.5 Success issues

Sometimes, a project's biggest problem is that it succeeds too well. A recent blockchain development in the Hyperledger environment went live and it exceeded its planned growth by 300% leading to performance and storage issues. The need for successful projects to have elastic storage (that is scalable and secure) is clear. If existing data stores are used, how will the success of a blockchain application that is parasitic on a data store affect the existing application and its performance? How will the influx of new data, possibly not linked in the same manner as far as foreign keys and other relationships affect the existing data store? These questions must be

addressed before they cause existing applications to fail or give incorrect results caused by the encroachment of the blockchain data.

Let's look at a simple calculation to see what the effect on storage volume of a successful blockchain project using offchain storage would entail. Our definition of success would be an increase in the transactions per second, so going from 1 to 10 to 100 or higher means the project is more successful. For this example, we will assume that every third transaction results in the generation of an offchain document. We will use an average document size, so the type of document doesn't matter for this example. We will also assume that the business is not a 24x7 operation open 364 days a year, however, it is a simple extrapolation should you iwsh to see the results form that type of operation.

Because non-transaction data, such as pictures, contracts, PDF, personal information and so on should not be stored in the actual blockchain, some form of off-chain or sideDB storage is required. Generally, off-chain data is unstructured. A hash or signature for the off-chain item is generated and that is what is stored in the blockchain. The actual item is either stored in the cloud or in near-cloud storage. It is expected that the required storage for off-chain data will exceed the needs of blockchain storage.

2.5.1 Off-chain data assumptions

The following assumptions are made in the off-chain storage calculations:

1. Each transaction does not produce a document. For calculation, a ratio of three transactions for each document is used so a figure of 0.3 is used.
2. 8 hours per day is the workday.
3. 240 days per year are actual processing days.
4. Indexing and other storage requirements are not included.

2.5.2 Calculation

Using the assumptions made in the previous section, the calculation becomes:

1 TPS*0.3 doc/transaction*3600 sec/hr*8 hr/day*240 day/yr=DPY (documents per year)
= 0.3*3600*8*240=814,301 DPY for 1 TPS

2.5.3 Storage required to save documents

According to current references, 1,000,000 documents (of mixed formats) requires 333 GB of storage. This means, 814,301 documents will require:

1000000X/(333*814301) = GB/year/TPS
X = (333*814,301)/1,000,000
= 271 GB/TPS/year = 0.264 TiB/year/TPS

So, as expected, you can see that if the transactions produce one document for every three transactions recorded, as the transactions per second increase, the storage required soon exceeds those required for the primary blockchain by a couple orders of magnitude.

2.6 New businesses

All new disruptive technologies bring new businesses. New businesses do not have existing data stores. So new storage is required for them to participate in the blockchain revolution.

2.7 New business processes require new data organization

Blockchain business application solutions often lead to re-engineering of business processes. For example, in *payments* (multi-industry), the business workflow might change to take advantage of the potential streamlining available in a blockchain community of participants. This could significantly transform an institution's payment systems, many of which have been in place for decades. Re-engineering may also be required to meet new performance standards, such as *near real-time* objectives. Most of the existing data stores are based (usually loosely) on the relational model of interrelated tables of data connected by primary and secondary key values. The tables are arranged in a way that the current application can quickly and easily access it. The blockchain requires a different form of data model which is more data warehouse-like than relational where the blockchain acts as the central *fact* table and the off-chain storage acts as the *dimension* table. The blockchain only stores the facts (totals, aggregations, transaction details), that result from or generate the intersection of data (documents, pictures, PDF files, tabular data) stored offchain. This new data organization requirement results in inefficiencies, need for re-organization, and changes to the way existing data stores store information.

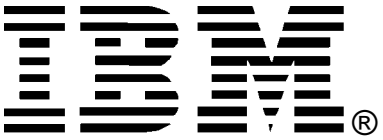
2.8 GDPR and granulation of data

General Data Protection Regulation (GDPR) in Europe and companies doing business in Europe will also drive the need for new off-chain storage in blockchain applications. It is recommended to store sensitive information offchain so that you can delete if need be.

3 The solution

The only way to fully, uniformly, and securely control access to off-chain data and mitigate all of the issues mentioned earlier, is to create a shared network of storage and server resources that is designed to provide the required security and shared environment for the blockchain consortium members. Each time a data object is accessed it must be verified using stored hash values proving that it is the same object as the one that was stored initially. Each object should be stored in more than one data store to make sure that loss of one node does not result in significant data loss as long as the node is down. Also, once a node rejoins the consortium after recover, a mechanism to synchronize the off-chain references and rebalance the off-chain data is also be required. These requirements and few others drive the need for new blockchain dedicated off-chain storage assets and new storage access paradigms.

The storage calculations show that unless the blockchain is being used for low volume, cryptocurrency-based transactions, it is recommended that a permissioned blockchain (such as IBM Hyperledger) with a fixed price per month be utilized.



IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PAPER “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes may be made periodically to the information herein; these changes may be incorporated in subsequent versions of the paper. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this paper at any time without notice.

Any references in this document to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
4205 South Miami Boulevard
Research Triangle Park, NC 27709 U.S.A.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

If you are viewing this information softcopy, the photographs and color illustrations may not appear.

Trademarks