

# Blockchain Technology – A Decentralized Solution To Data Deduplication In Cloud Storage Systems

Padma Punna<sup>1</sup> , Dr.B.Tirapathi Reddy<sup>2</sup> , Yada Sunitha<sup>3</sup>

<sup>1</sup>Research scholar, Department of Computer Science, KL University (A.P) & DL in  
Telangana Tribal Welfare Residential Educational Institutions , Hyderabad.

<sup>2</sup>Professor & HOD, Department of Artificial intelligence & Data Science, KL University  
(A.P)

<sup>3</sup>Assistant Professor , Dept. of CSE , CVR College of Engineering Hyderabad.

---

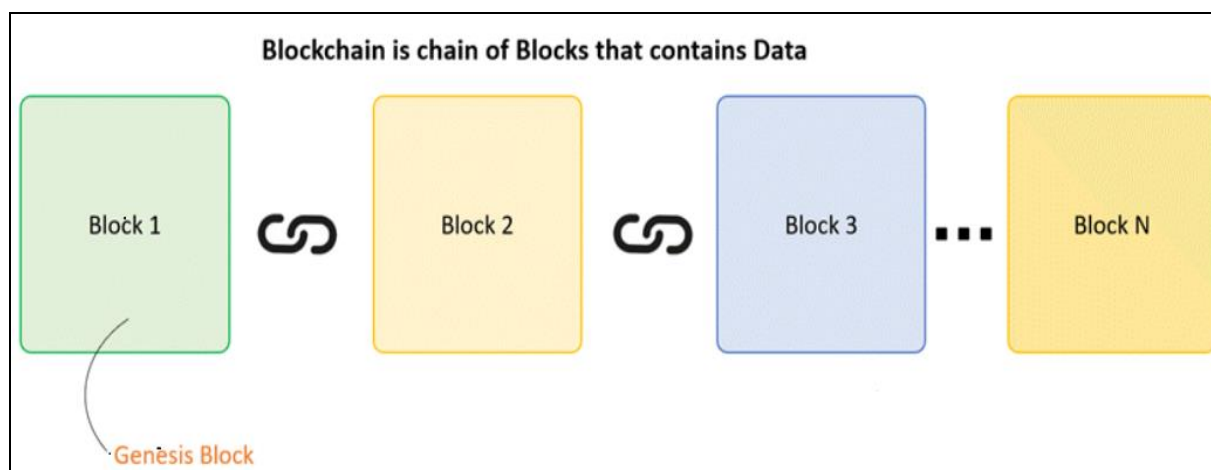
## Abstract:

Cloud computing used by those who are unable to maintain their own infrastructure, security features, or storage system capabilities. Cloud computing is the ideal alternative to on-premises computing since it allows users to pay only for the storage or services that they really utilise. Cloud computing comes with its own set of problems to overcome. In Cloud computing, service providers deliver services to several consumers at the same time. Because it is maintained by numerous customers, duplicate data on the cloud platform slows down the cloud environment's performance. With deduplication, you may get away of irrelevant stuff from your storage system and boost its speed. During this paper we have a tendency to gift a comprehensive summary on Blockchain technology and how it is useful in removing data deduplication. Blockchain Technology could be absolutely localised digital register that keeps a secure history of information exchanges. The decentralization face of Blockchain Technology will away the necessity of any central authority for managing it. This authorisation of uploaded data may be achieved utilising a decentralised system known as Blockchain, which allows one to authorise data uploads and claim ownership of the material, therefore granting the user's copyrights. Data deduplication is combined with Blockchain in this article, resulting in picture storage that is more efficient, effective, and secure.

**Keywords** - BlockChain, Data Deduplication

## 1. Introduction to BlockChain Technology

Blockchain is a decentralised, distributed network that offers immutability, privacy, security, and transparency. Decentralization, persistency, anonymity, and auditability are the four main properties of the Blockchain.



Blockchain is a distributed ledger technology[2] that is most often used in the Bitcoin cryptocurrency. According to the Financial Times (2016), Blockchain is a "network of computers, all of which must authorise a transaction before it is recorded in a 'chain' of coding system." The key elements of the transfer are recorded in a public ledger that is visible to everyone on the network."

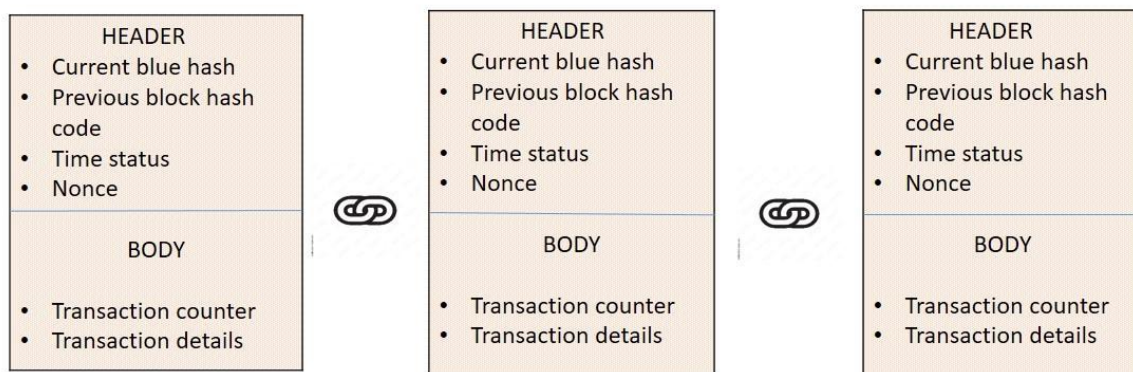
Satoshi Nakamoto published a white paper titled "Bitcoin: A Peer-to-Peer Electronic Cash System" in 2008, which gave birth to blockchain technology.

Blockchain technology is a peer-to-peer distributed structure that could be used to solve the problem of data duplication in cloud storage systems[1]. It is a collection of blocks that contain encrypted transactional details and share the same timestamp. The network nodes (miners) are in charge of connecting the blocks in sequential order, with each block containing the hash of the previous block eventually serving as the digital signature for each block. The digital signature consists of and are determined by two variables: one that contains transactional information, and the other that stores hash value of the previous block within the chain.

To accomplish this, famous multiple hashing algorithms such as SHA256 and RSA can be used.

This paper aims to provide a solution to data duplication in Cloud Storage Systems using blockchain that is comparable to existing methodologies. Additionally, it will also enlighten upon the restrictions of blockchain implementation as well as the future scope.

### 1.1 Structure of a Block in BlockChain



The genesis block, which has no parent block, is the first block in a blockchain.

#### 1.2 Block Header:

A block header [11] is employed to spot a specific block on a whole blockchain. Every of the blocks contains a novel header, and every such block is identified by its block header hash individually. Header contains below mentioned fields.

#### 1.3 Blockchain Version:

It's a four-byte parameter that indicates the set of block validation criteria that must be followed. As there are four types of Blockchain versions, this parameter helps in identifying the version that the current block is using.

#### 1.4 Version 1.0 :

This version is focussed on Cryptocurrency and employs a public ledger, like as Bitcoin, to store the information.

#### 1.5 Version 2.0:

This version referred to as smart contracts, which are self-executing programmes, such as Ethereum.

#### 1.6 Version 3.0 :

It is not intended to create a decentralised structure, the best example for this Tor Browser.

#### 1.7 Version 4.0 :

This version is primarily used for Industry to establish a scalable and, cheap network that can be utilised by many people.

#### 1.8 Merkle tree root hash:

At block intervals, the hash value of all transactions. A Merkle root visualises if data has been tampered with, hacked, or modified using mathematical methods. The Merkle Root is a hash of a block that is calculated using the binary tree concept.

#### *1.9 Timestamp:*

The timestamp is used as proof that a specific block was used at a specific moment. The timestamp helps in validating the authenticity of a block.

#### *1.10nBits :*

It's 32 bit target threshold of a legitimate block hash worth . nBits cannot be modified in the slightest degree. It's worth should equal the (compact encoding of) the target worth for that block, that is decided entirely by the predecessors of the block.

#### *1.11Nonce:*

It's short form for “number only used once” and it describes the average number of miners found. A nonce might be a 32-bit number with a maximum value of  $2^{32}$  potential possibilities.

#### *1.12Parent block hash :*

This describes the previous block's 256-bit hash. The prior hash holds the hashed value of the preceding node's address, while the Genesis Block is the first block in the blockchain and has no previous block hash value.

#### *1.13 Consensus Algorithms*

A consensus algorithm [10] is a process through which all peers of Blockchain nodes come to a consensus on the current state of the distributed ledger. In a distributed computing system, consensus methods ensure stability in the Blockchain network and build confidence amongst unknown peers.

The consensus mechanism ensures that every block added to the Blockchain is the one and only version of the truth that all nodes in the blockchain agree on.

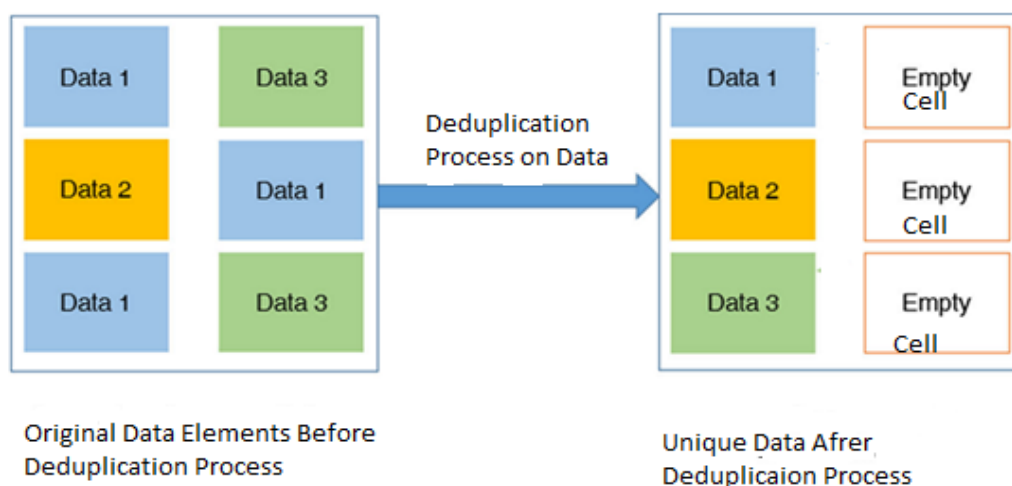
The Blockchain consensus algorithm has several specific goals, including reaching a common accord, collaboration, co-operation, equal rights for all nodes, and requiring each node to participate in the consensus process. A consensus algorithm seeks to identify a common accord that benefits the entire network in this way.

## **2 Data deduplication in Cloud Storage System**

For users who don't have access to their own comprehensive infrastructure, or who don't know how to maintain security features and storage systems, cloud computing is a viable option to consider. In order to pay just for what they really use, cloud computing is a great option for all of these people. However, cloud computing has its own set of issues. Multiple customers may

utilise the cloud infrastructure at the same time, which is provided by cloud service providers. More than half of all firms have moved their data to cloud storage by 2020, according to cloud market stats cited in this article.

Cloud storage providers[5] are seeing a significant increase in the amount of data they can store. In the event of a data loss, distributed storage is utilised to provide high dependability, higher availability, and disaster recovery. The number of cloud storage replications has an impact on the system's performance. The replication factor refers to the bare minimum of duplicate data. This is critical to the storage system's dependability and availability. Conversely, duplicate data puts additional strain on the storage system, requiring more space and bandwidth to handle them. Deduplication is the storage system technology used to decrease this duplication. High-level data deduplication is shown in below figure.



Deduplication is the process to identify and deleting redundant data from the storage system, and then storing the unique data in the Cloud Storage system.

Benefits of the Deduplication:

- Increases the storage capacity of the system Improves the system's usability and accessibility.
- Increases the storage unit's bandwidth
- Lowers the system's cost as well as the cost of maintenance.

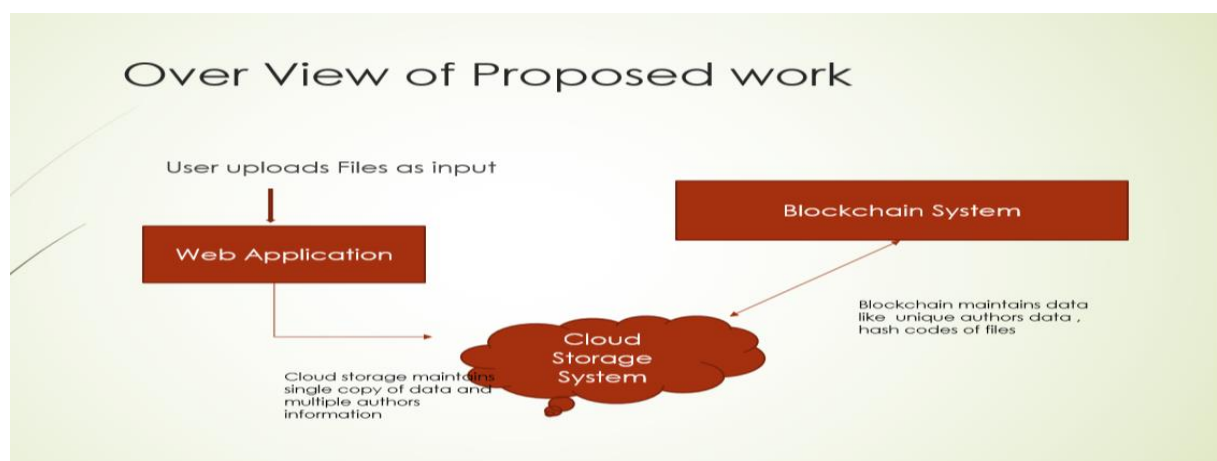
### 3 Existing Works in Data Deduplication by using blockchain technology

Sno	Author Name	Paper Title	Proposed Work
-----	-------------	-------------	---------------

1	SHANGPING WANG <sup>1</sup> , YUYING WANG <sup>2</sup> , and YALING ZHANG <sup>3</sup>	Blockchain-based fair payment protocol for deduplication cloud storage system, IEEE paper published in September 2020	Proposed work is, For cloud storage systems, a fair payment scheme based on the Ethereum blockchain is suggested. The blockchain technology enables decentralised payment, and the fairness of payment is ensured by a smart contract with a pre-existing penalty. In this system, there is no trustworthy third party. The payment procedure is decentralised thanks to the use of ethereum smart contract technology to transmit tokens over the blockchain network. On the one hand, deduplication technology is used to provide effective and secure methods for cloud storage servers and clients to reduce storage, communication, and computing overhead; on the other hand, deduplication technology provides two different payables, for new file one payment and another is "whether files uploaded by the client are duplicated with existing files on the cloud storage server" another payment.
2	Jingyi Li, Jigang Wu, Long Chen, Jiaying Li	Secure and Reliable Distributed Deduplication with Blockchain	To provide efficient deduplication with great reliability, a novel blockchain-based distributed deduplication system is presented. File tags are stored on the blockchain and are used to check for duplicates and data integrity locally. Our architecture, which combines a traditional encryption technique with an incentive mechanism, can perform online missing file restoration via an automatically executed smart contract. • The proposed construction is efficient, and the

			redundancies are optimised and equivalent to other storage systems providing the same degree of dependability, according to the evaluation results.
3	Aparna R., Roopa G. Kulkarni, and Shilpa Chaudhari	Secure Deduplication for Images using Blockchain, IEEE Paper 2020	Implemented Web Application, which takes input image from the user along with Owner information. Check the deduplication by using SHA 256, if file is unique it will move to cloud and user data maintained in blockchain. If file is duplicated it will not uploaded.
4	B. Tirapathi Reddy , M. V. P. Chandra Sekhara Rao	Privacy preserving proof of ownership for data in cloud storage systems	Ex OR (fingerprint hash, File hash) is checked for by the proposed Cooekoo filter, which assumes role-based ownership of the data algorithm. It establishes file ownership and gives confirmation of ownership of material that is being uploaded to the cloud.
5	B. Tirapathi Reddy , M. V. P. Chandra Sekhara Rao	Filter Based Data Deduplication in Cloud Storage using Dynamic Perfect Hash Functions	This paper proposed Cooekoo based Filter algorithm to upload the files in to the CSP . This Cooekoo filter eliminates the duplicated files, so redundant data can be upload to the CSP . Compared to Bloom Filter and D-Left Bloom Filter Cooekoo filter gave better results in terms of filter occupancy and execution time. Cooekoo filter guarantees the maximum elimination of duplicated data

#### 4 Proposed System



Blockchain technology image/file deduplication is a web-based application. For those who want to keep their files in the cloud and assert ownership, this programme provides a user interface. Users are prompted to provide their personal information on the website's homepage before uploading images to a cloud storage platform. The handler or controller in the MVC framework receives and processes requests from the UI page. The hash value of the image to be uploaded is included in the message digest object that is created by the controller. This information is only stored in the local MYSQL database for a given image if it was previously inaccessible in the database. If already file is available in storage system, system will accept the file with authenticated data and file owner will be added to the existing authors. These one-of-a-kind files are stored in the cloud using one of the AWS or Azure services. Blockchain will be maintained with the unique files / image and it will be immutable forever. Below is a flow diagram for the proposed work.

#### Flow Chart of Proposed System





### Advantages of Proposed System

Proposed system Deduplication by Blockchain technology is removes duplication and also

- No Third Party interference
- Enhanced security
- Greater transparency
- Instant traceability
- Increased efficiency and speed

### 5 Conclusion

Blockchain uses peer-to-peer technology to store and transfer data. It's decentralised and doesn't need the participation of a third party to work. Peers may exchange data without the need for intermediaries or other trusted third parties using blockchain technology. It functions in a decentralised structure that is governed by the rules established by the network's members. Mainly focused about how to use block chain technology to remove data duplication in cloud storage system. Considering the benefits and drawbacks of the suggested web-based

application is an important first step in my work. We have also outlined its proposed application and the feasibility of the system. The aim of this paper is to provide a comprehensive readymade idea of the removing data deduplication in cloud storage system by using blockchain technology.

#### **Declarations :**

##### **1. Funding Details :**

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

##### **2. Conflicts of interest/Competing interests:**

We declare that the all the authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

##### **3. Data availability statement:**

All authors declare that the results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration (from you or one of your Contributing Authors) by another publisher.

##### **4. Author contribution:**

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Padma Punna. The first draft of the manuscript was written by Padma Punna and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

#### **6 References**

1. Nikita Rajeshkumar Bagrecha, Ishaq Mustafa Polishwala, Pragya Abhai Mehrotra, Rishabh Sharma, B.S. Thakare , "Decentralised Blockchain Technology: Application in Banking Sector", 2020 International Conference for Emerging Technology (INCET), 2020
2. Tareq Ahram, Arman Sargolzaei, Saman Sargolzaei, Jeff Daniels, Ben Amaba. "Blockchain technology innovations", 2017 IEEE Technology & Engineering Management Conference (TEMSCON), 2017
3. Arjun Kumar, Adesh Sangoi, Shubham Raj, Kiran M. "ShardCons - A Sharding Based Consensus Algorithm for Blockchain", 2021 IEEE International Conference "Blockchain Applications in IoT Ecosystem", Springer Science and Business Media LLC, 2021
4. Aparna R., Roopa G. Kulkarni, and Shilpa Chaudhari , "Secure Deduplication for Images using Blockchain", IEEE Paper 2020
5. Jingyi Li, Jigang Wu, Long Chen, Jiaying Li, "Secure and Reliable Distributed Deduplication with Blockchain" Lecture Notes in Computer Science book series (LNTCS, volume 11334)

6. B. Tirapathi Reddy , M. V. P. Chandra Sekhara Rao "Privacy preserving proof of ownership for data in cloud storage systems" published in March 2018 International Journal of Engineering & Technology
7. SHANGPING WANG<sup>1</sup> , YUYING WANG<sup>2</sup> , and YALING ZHANG<sup>3</sup> "Blockchain-based fair payment protocol for deduplication cloud storage system" IEEE Paper 2020
8. B. Tirapathi Reddy , M. V. P. Chandra Sekhara Rao "Filter Based Data Deduplication in Cloud Storage using Dynamic Perfect Hash Functions" DOI 10.5013/IJSSST.a.19.04.08
9. <https://www.geeksforgeeks.org/consensus-algorithms-in-blockchain/#:~:text=A%20consensus%20algorithm%20is%20a,state%20of%20the%20distributed%20ledger.>
10. <https://www.geeksforgeeks.org/different-version-of-blockchain/>