Data download Duplication Alert System

Sanjana K R[1], Sinchana M[2], Sindhu T[3], Shweta Salimath[4], Sheethal S[5]
[1,2,3,4,5]*Department of CSE, Reva University, Bangalore, India*

***Abstract*—Organizations often face data redundancy when multiple users download the same datasets without awareness of existing copies, resulting in wasted storage, bandwidth, and time. The Data Download Duplication Alert System (DDAS) addresses this issue by detecting and preventing duplicate dataset downloads using content-based identification. The system maintains metadata logs containing file details such as content identifiers, ownership, and timestamps. IPFS-based hashing enables accurate detection of duplicate datasets even when file names differ, while Firebase provides real-time synchronization and alerting. A React-based interface allows users to efficiently search and manage datasets. By integrating decentralized storage with real-time metadata management, DDAS improves resource efficiency and supports collaborative data usage.**

***Index Terms*— data deduplication, IPFS, Dataset Management, Firebase, Content Identifier, Decentralized Storage.**

I. INTRODUCTION

In today's data-driven environment, organizations and institutions work with large volumes of data every day. Researchers and professionals often need access to similar datasets for analysis, experiments, or reporting. However, due to limited visibility into previously downloaded data, users may unknowingly download the same dataset multiple times. This results in unnecessary use of storage space and bandwidth, along with added complexity in managing data efficiently. Traditional centralized storage systems face challenges such as redundancy and inefficiency in collaborative environments [5], [6].

The Data Download Duplication Alert System (DDAS) is designed to solve this problem by automatically identifying and preventing duplicate dataset downloads. The system keeps a detailed record of all downloaded files, including important information such as file name, owner, department, content identifier (CID), and time of download. Whenever a user attempts to download a dataset, the system checks the file's IPFS hash to see if the same content already exists in the system. If a duplicate is detected, the user is alerted and redirected to the existing dataset instead of downloading it again.

By combining IPFS for content-based identification, Firebase for real-time updates, and a simple React-based interface, DDAS offers an efficient and user-friendly solution for data management. The system helps reduce resource wastage, improves transparency, and encourages better collaboration by allowing users to reuse existing datasets rather than creating unnecessary duplicates.

II. DATASET STORAGE MECHANISMS

Dataset storage systems are generally categorized into centralized and distributed architectures based on how and where data is stored. Centralized storage systems are the most commonly used approach in organizations and academic institutions. In this model, datasets are stored on a single server or within a controlled cloud environment, such as institutional data centers or commercial cloud platforms. These systems are easy to manage, provide controlled access, and allow administrators to monitor data usage from a central location.

Despite these advantages, centralized storage systems have several inherent limitations. Since all data is stored in one place, the system is vulnerable to a single point of failure. Any hardware malfunction, network outage, or security breach can result in data unavailability or loss. As the volume of data and number of users increase, centralized systems may also face scalability challenges, leading to performance degradation and higher maintenance costs. Additionally, access to datasets is often dependent on continuous network availability, which can restrict usability in distributed or remote research environments.

A significant drawback of centralized storage systems is their lack of effective duplicate detection mechanisms. Centralized storage systems often fail to prevent duplicate datasets due to reliance on metadata rather than content-based identification [5]. Most centralized platforms rely on file names or user-provided metadata to manage datasets. As a result, identical datasets may be stored multiple times under different names or descriptions, making redundancy difficult to detect. This issue becomes more pronounced in collaborative research settings, where multiple users independently download or upload the same dataset without awareness of existing copies.

Such redundancy leads to inefficient use of storage space, unnecessary bandwidth consumption, and increased data management complexity. Moreover, the absence of automated duplicate verification increases the chances of repeated research efforts, as users may unknowingly work on datasets that already exist within the system. These limitations highlight the need for smarter storage solutions that can identify datasets based on their actual content rather than surface-level attributes.

III. DECENTRALIZED STORAGE SYSTEMS

Decentralized storage systems offer an alternative to traditional centralized storage by distributing data across multiple nodes instead of relying on a single server. This approach reduces the risk of system failure and improves data availability, as files can be accessed from multiple locations. By removing dependence on a central authority, decentralized systems provide better resilience, fault tolerance, and reliability, especially in data-intensive environments

One of the most widely used decentralized storage technologies is the Interplanetary File System (IPFS). IPFS follows a content-based addressing model, where files are identified using cryptographic hashes known as Content Identifiers (CIDs) [1]. Instead of referring to files by their location or name, IPFS identifies data based on its actual content. This means that any change in the file, even a minor one, results in a different CID.

A key advantage of IPFS is that identical files always produce the same CID, regardless of file name or where the file is stored. This property makes IPFS particularly effective for detecting duplicate data and ensuring data integrity. Several studies have shown that IPFS improves data availability and security by distributing content across multiple peers, reducing the risk of data loss and unauthorized modification when compared to traditional centralized storage systems.

Blockchain-based storage solutions have also been studied as part of decentralized data management [2]. While blockchains provide strong immutability and transparency, storing large datasets directly on blockchain networks is not practical due to high storage requirements and transaction costs. To overcome this limitation, many modern systems adopt a hybrid approach, where IPFS is used to store large files and an external database is used to manage metadata and access information. This combination balances decentralization with performance and usability.

IV. SURVEY OF DEDUPLICATION TECHNIQUES

Data deduplication techniques are used to eliminate redundant copies of data in storage systems by identifying identical content. Traditional deduplication methods are commonly implemented in centralized storage environments and rely on techniques such as file-level hashing, block-level comparison, or checksum verification. These methods compare stored data blocks to detect redundancy and store only a single copy of identical data.

While effective in reducing storage usage, these approaches are usually applied at the backend and are not visible to end users. As a result, users may still upload or download duplicate datasets without being notified. Additionally, conventional deduplication systems often require high computational resources to compare large datasets and may not scale efficiently in collaborative environments. Existing deduplication techniques primarily operate at the backend storage level and do not provide user-level duplicate alerts [5].

In research and academic settings, where multiple users independently access datasets, traditional deduplication techniques fail to provide real-time feedback or alerts. This limitation highlights the need for a deduplication mechanism that operates at the user interaction level and is capable of identifying duplicate content before unnecessary data transfers occur.

V. DECENTRALIZED STORAGE USING IPFS

The Interplanetary File System (IPFS) introduces a decentralized approach to file storage by addressing data based on its content rather than its location. Each file stored in IPFS is assigned a unique cryptographic hash called a Content Identifier (CID), which represents the exact content of the file. If two files have identical content, they will generate the same CID, regardless of their file names or sources.

This content-based identification makes IPFS inherently suitable for duplicate detection. Unlike traditional storage systems, IPFS does not store multiple copies of the same file, thereby reducing redundancy and improving storage efficiency. Additionally, the decentralized nature of IPFS enhances data availability and fault tolerance, as files are distributed across multiple nodes rather than stored on a single server.

IPFS has gained attention in research literature due to its ability to ensure data integrity, reduce dependency on centralized infrastructure, and support scalable data sharing. These characteristics make it a strong candidate for applications that require reliable duplicate detection and decentralized data management.

VI. COMPARATIVE ANALYSIS AND RESEARCH GAP

A comparative analysis of existing storage and deduplication systems reveals several gaps in current approaches. Centralized storage systems provide ease of management but lack effective content-based duplicate detection and suffer from scalability and reliability issues. Traditional deduplication techniques reduce storage redundancy but are typically hidden from users and do not prevent duplicate downloads at the interaction level.

Decentralized storage systems such as IPFS address many limitations of centralized storage but often focus primarily on file storage without sufficient integration of metadata management and user-level coordination. Many existing systems fail to combine decentralized content identification with real-time collaboration and alert mechanisms.

This analysis indicates a clear research gap: the absence of a unified system that integrates decentralized storage, content-based duplicate detection, and real-time metadata tracking to prevent redundant dataset usage in collaborative environments.

VII. WHY THE DDAS APPROACH IS NEEDED ?

The Data Download Duplication Alert System (DDAS) is proposed to address the limitations identified in existing systems. By leveraging IPFS for content-based identification, DDAS ensures accurate detection of duplicate datasets using CIDs. At the same time, real-time metadata management allows users to view existing datasets and receive alerts before downloading duplicates.

Unlike traditional deduplication systems, DDAS operates at the user interaction level, preventing redundancy before it occurs rather than after storage is consumed. The integration of decentralized storage with real-time synchronization improves transparency, resource utilization, and collaboration across departments.

Overall, the DDAS approach provides a practical and efficient solution for modern data-driven environments, where preventing redundancy, conserving resources, and enabling collaborative data sharing are critical requirements.

VIII. METHODOLOGY

The proposed Data Download Duplication Alert System (DDAS) will follow a structured methodology aimed at preventing redundant dataset downloads within an organization. The system will be designed using a modular architecture that integrates decentralized storage, real-time metadata management, and a user-friendly interface.

Initially, the system will require users to authenticate themselves through a secure login mechanism. Once authenticated, users will be able to upload or request datasets through a web-based interface developed using React [4]. When a dataset is selected, the system will generate a cryptographic hash using the Interplanetary File System (IPFS) [1], producing a unique Content Identifier (CID) that represents the actual content of the file.

This generated CID will then be compared against existing CIDs stored in a centralized metadata repository managed using Firebase [3]. If a matching CID is found, the system will identify the dataset as a

duplicate and alert the user, providing access to the already available dataset. If no match is found, the metadata associated with the dataset such as file name, owner, department, and timestamp will be stored for future reference.

Fig. 1. This methodology ensures that duplicate detection will occur before unnecessary data downloads take place, thereby reducing storage overhead and bandwidth usage.

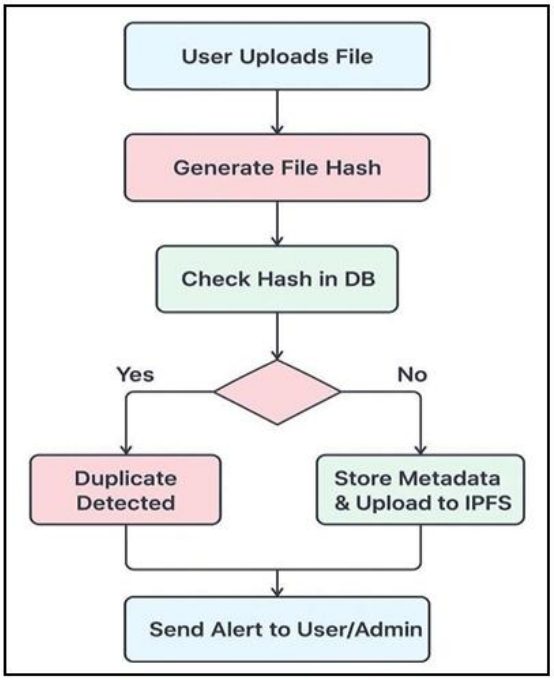

Fig. 1. Proposed System Workflow Diagram

IX. RESULTS

The proposed DDAS is expected to significantly reduce redundant dataset downloads by identifying duplicates based on content rather than file names. By leveraging IPFS hashing, the system is expected to detect identical datasets accurately even when they are uploaded or requested under different names.

Real-time metadata synchronization using Firebase is expected to improve visibility of existing datasets across users and departments. As a result, users will be informed about previously downloaded datasets, reducing repeated downloads and improving coordination.

The system is also expected to optimize resource utilization by conserving storage space and minimizing unnecessary bandwidth consumption. Overall, DDAS is anticipated to provide a more efficient and transparent dataset management environment compared to traditional centralized systems.

Tablel Expected Benefits of the Proposed DDAS System

Aspect	Expected Outcome	Benefit
Duplicate Detection	Identical datasets will be identified using IPFS-based Content Identifiers (CIDs)	Prevents repeated downloads of the same dataset
Storage Utilization	Reduced storage usage by avoiding redundant copies of datasets	Efficient use of storage resources
Bandwidth Consumption	Fewer unnecessary data transfers across the network	Saves network bandwidth and reduces load
Data Visibility	Users will be informed about already available datasets	Improves transparency and awareness
Collaboration	Users can reuse existing datasets instead of downloading again	Encourages collaboration and shared data usage
Data Management	Centralized metadata tracking with decentralized file identification	Simplifies dataset tracking and organization

X. DISCUSSION

The proposed approach addresses key limitations identified in existing dataset storage and deduplication systems. Unlike traditional centralized storage systems that lack content-based duplicate detection, DDAS will use CID-based identification to ensure accurate and reliable detection of duplicate datasets.

By operating at the user interaction level, the system will prevent redundancy before it occurs, rather than attempting to eliminate duplicates after storage resources have already been consumed. The integration of decentralized content identification with real-time metadata management is expected to enhance transparency and promote collaborative data sharing.

Although the system focuses primarily on duplicate detection and alerting, the proposed architecture

allows for future enhancements such as access control mechanisms, usage analytics, and integration with institutional data governance policies. Overall, DDAS is expected to provide a practical and scalable solution for organizations seeking to improve data efficiency and coordination.

XI. CONCLUSION

The Data Download Duplication Alert System (DDAS) addresses an important yet often overlooked issue in organizational data management, namely the repeated downloading of identical datasets. By combining decentralized content identification using IPFS with real-time metadata synchronization through Firebase and a user-friendly React-based interface, the proposed system offers a practical approach to detecting duplicate datasets and alerting users before unnecessary downloads occur. This helps in conserving storage space, reducing bandwidth usage, and improving overall data visibility within an organization.

The use of content-based identification ensures accurate duplicate detection even when dataset names differ, while real-time updates enable users to stay informed about existing datasets. As a result, DDAS encourages collaboration, transparency, and reuse of available data rather than redundant downloads. The proposed approach is adaptable to various domains, including academic institutions, research organizations, and enterprise environments. Future extensions of the system may include advanced analytics, intelligent dataset classification, and enhanced access control mechanisms. Overall, DDAS presents a scalable and efficient solution for improving data management and resource utilization in collaborative digital environments.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Prof. Shwetha Salimath, whose invaluable guidance and insightful feedback shaped this project from its initial concept to its final form. My sincere thanks go to my colleagues, Sinchana M, Sindhu T and Sheethal S for their collaboration, commitment, and contributions to this research. I gratefully acknowledge the support of Reva University for providing the resources and facilities which made this work possible. Finally, I am deeply thankful to my family and friends for their unwavering support and encouragement, which has been invaluable throughout this journey.

REFERENCES

[1] J. Benet, “IPFS – Content addressed, versioned, P2P file system,” 2014.

[2] G. Wood, “Ethereum: A secure decentralised generalised transaction ledger,” 2015.

[3] “Firebase documentation – Authentication and Firestore,” *Firebase*, [online]. Available: https://firebase.google.com/docs

[4] “React.js official documentation,” *React*, [online]. Available: https://react.dev

[5] R. Kaur and S. Khanna, “A review of data deduplication techniques for storage optimization,” *Int. J. Comput. Appl.*, 2021.

[6] S. Wilkinson *et al.*, “A survey on decentralized storage systems,” *IEEE Acces*