**Secure Hashing Techniques with CRC32b for Efficient Data De-duplication**

**Abstract –**

The exponential growth of digital data in storage systems is a critical issue presently as a large amount of duplicate data in the storage systems exerts an extra load on it. De-duplication is a useful technology that is gaining popularity in large-scale storage systems. De-duplication removes duplicate data, increases storage usage, and lowers storage costs. It examines modern data management issues, which are mostly caused by an excess of digital data and data redundancy concerns. The main goal is to enhance data management and storage by carefully detecting and eliminating duplicate files using the CRC32b safe hashing technique. This approach generates unique hash values for each data component, enabling precise identification and elimination of duplicate files while retaining data security and integrity. The study evaluates the CRC32b secured hashing method's appropriateness for data de-duplication in a variety of usage situations. It contains a thorough examination of the CRC32b method as well as the creation of a data de-duplication system capable of handling a variety of scenarios such as personal data management, cloud storage, and enterprise-level data centers. The system's performance is evaluated using quantitative metrics such as time spent processing, resource consumption, and disk space savings. This study provides substantial information into data redundancy, storage inefficiency, data integrity, and security, offering a full understanding of these complex subjects. The use of CRC32b secure hashing technology for data de-duplication is revolutionary and offers a new way to dealing with data redundancy. Because of this broad reach, the findings are broadly relevant and advantageous to a variety of data management scenarios.

1. **Introduction**

The development of cloud computing has caused a notable rise in data quantities in the age of digital transformation. The popularity of social networking sites, the expanding use of smartphones, and the expansion of the internet are the main drivers of this upsurge. Businesses all throughout the world are finding it difficult to store and handle these enormous amounts of data. Data are frequently copied across several storage sites to improve availability, dependability, and disaster recovery. Nevertheless, the storage system is burdened further by this duplication, which necessitates more bandwidth and storage capacity. A unique kind of data compression called data de-duplication has become a vital piece of technology for effective large data storage. By getting rid of unnecessary data, cloud storage systems' storage capacity and network transmission rate are decreased. Duplication compensates for the increasing demand for storage capacity by detecting duplicate data, preserving just one copy, and strategically employing logical pointers for duplicated data. Data de-duplication techniques are increasingly widely used by cloud storage providers and backup services to increase storage efficiency (Kaur et al.).

However, there are differences in the de-duplication methods used for text, picture, and video content, as these methods depend on the kind of data being used. This kind of information is important for the creation of de-duplication strategies since these data types have various implicit features and storing formats. The effectiveness and efficacy of data matching algorithms pose a barrier to de-duplication approaches for big distributed storage systems, despite the industry's enormous benefits. The rise in the usage of cloud computing, attributed to its lower operational costs, scalability, and easier access, has paved the way for wider use of cloud storage services. These services not only protect user data from local hazards but also improve portability, facilitating data sharing among different users and enabling simultaneous work on the same project. As researchers in academia and industry strive to develop efficient distributed de-duplication techniques, this paper delves into the intricacies of secure hashing techniques with CRC32b for efficient data de-duplication (Moorthy et al.).

1. **Context of Data Duplication Problem**

Data duplication poses a significant challenge in the contemporary digital landscape, as it can detrimentally affect system performance, storage capacity, and data retrieval processes. Various factors, including user activities, system backup procedures, and data transfer protocols, can inadvertently lead to the creation of duplicate files. The unintentional generation of duplicate data instances not only squanders valuable storage resources but also introduces inefficiencies into the data management process. This issue is particularly pronounced in large-scale data storage environments where the frequency of data duplication can result in substantial financial and operational costs. It is therefore imperative to address this issue with effective strategies and technologies.

Chart, bubble chart

Description automatically generated

Figure 1: Data De-Duplication for Two files split in chunks

As shown in Figure 1, data de-duplication is a procedure that successfully reduces duplicate data segments to unique segments. This is accomplished by segmenting the whole file into fixed or variable-size segments. The essential idea of the de-duplication process is to store just one duplicate of each segment. When the de-duplication engine meets data that already exists in the storage system, it does not store another duplicate of the same data. Instead, it constructs a pointer at the place of the duplicate data that points back to the original copy. This strategy is useful for freeing up blocks inside the storage system, hence optimizing memory space. Figure 1 depicts the de-duplication process clearly, exhibiting how duplicate segments are reduced to unique segments, resulting in more efficient use of storage resources.

A major issue occurs in the big data landscape within a cloud environment—duplicate data. As the internet's popularity expands, so does the massive digital footprint, necessitating a greater dependence on big data. However, the daily influx of unstructured data, comprising both pertinent and irrelevant information, poses a challenge in effective data management. Cloud computing technology emerges as a favored solution for companies grappling with this data deluge. Yet, the data collected from diverse sources often carries the burden of duplications, adding to the overall data volume without a corresponding increase in meaningful insights. To tackle this problem, a crucial strategy comes into play—the de-duplication process. This process becomes the linchpin for data handlers aiming to streamline their datasets by removing redundancy. The goal is clear: enhance data reliability and quality. This discussion zooms in on the intricacies of addressing duplicity within big data, with a specific focus on the unique challenges posed by a cloud storage environment.

In today's digital landscape, the challenge of data duplication has become a significant concern, prompting organizations and individuals to leverage cloud storage providers for data management. While cloud storage providers implement de-duplication techniques to optimize storage space by avoiding redundant data copies, a critical issue arises when subscribers encrypt their data for security before outsourcing it to the cloud. The encryption of data introduces a dilemma: de-duplication cannot be applied to encrypted data due to the generation of distinct ciphertexts using different encryption keys from various subscribers. This limitation poses a security-versus-efficiency conundrum, leading to the exploration of secure de-duplication methods. The paper underlines the importance of reviewing state-of-the-art methods addressing this challenging task of securely performing de-duplication over encrypted data in the cloud. Despite the advantages of cloud storage, such as scalability, cost-efficiency, and accessibility, the inability to distinguish identical encrypted data becomes a hurdle for Cloud Service Providers (CSPs). Typically, CSPs store a single copy of identical data from multiple sources to optimize space, but encrypted uploads with different keys thwart their ability to discern identical content. The rise in cloud computing adoption facilitates the storage of vast amounts of data, fostering data sharing and collaborative work on shared projects. However, the data collected from various sources often contains duplications, increasing data size without enhancing meaningfulness. Herein lies the significance of the de-duplication process in big data handling, serving as a vital solution to expunge redundancies and enhance data reliability and quality. In essence, the complex interplay of data security, storage optimization, and cloud computing dynamics underscores the data duplication problem. Striking a delicate balance between encryption for security and de-duplication for storage efficiency emerges as a central challenge in this multifaceted landscape, where the pursuit of data integrity coexists with the imperative for optimal storage practices (Mgwalima and Ndlovu).

1. **Research Questions**

The effective administration and preservation of digital data in the face of an increasing wave of redundancy and duplication is the research subject at the center of this dissertation. Data is created and saved in the digital age at a never-before-seen rate, requiring sophisticated methods to guarantee efficient data management, storage resource optimization, and quick data retrieval. Proliferation of duplicate files is one of the primary issues in this area and a major hindrance to data management systems in many sectors.

The primary research problems are framed as follows:

* 1. How does de-duplication technique enhance the performance of a large storage system and what are its pros and cons?
  2. How does the scheme support offline data sharing with reduplication without intruding the privacy of data holders?
  3. How are de-duplication techniques categorized based on storage, application point, and level?
  4. What are the potential future research directions in the field of de-duplication for academia and industry?

1. **Research Objectives**
2. The discourse has explored the significance of employing de-duplication techniques to enhance the efficiency of expansive storage systems. The examination encompasses an analysis of the essentiality of de-duplication techniques, along with an exploration of their advantages and drawbacks.
3. This solution displays a flexible capacity to promote data exchange via reduplication, even while the data bearer is offline. Importantly, it does this without jeopardizing the data owner' privacy.
4. Current de-duplication approaches are divided into three categories: storage-based, point-of-application-based, and level-based.
5. The dissertation emphasizes potential research pathways in the realm of de-duplication, providing as a guide for both academic and industrial researchers.

**References –**

* Kaur, Ravneet, et al. “Data De-duplication Techniques for Efficient Cloud Storage Management: A Systematic Review.” The Journal of Supercomputing, vol. 74, no. 5, 20 Dec. 2017, pp. 2035–2085, <https://doi.org/10.1007/s11227-017-2210-8>.
* Moorthy, Vaishnavi, et al. DE-DUPLICATION in CLOUD STORAGE USING HASHING TECHNIQUE for ENCRYPTED DATA. Vol. 13, no. 5, 2018, www.arpnjournals.org/jeas/research\_papers/rp\_2018/jeas\_0318\_6881.pdf. Accessed 25 Nov. 2023.
* Mgwalima, Tamanda, and Arthur Ndlovu. “A Review of the Past, Present, and Future of Secure Data De-duplication.” International Journal of Engineering Research & Technology, vol. 11, no. 12, 12 Dec. 2022, www.ijert.org/a-review-of-the-past-present-and-future-of-secure-data-de-duplication, https://doi.org/10.17577/IJERTV11IS120020. Accessed 25 Nov. 2023.