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

1. 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. Deduplication is an efficient technique that has gained attention in large-scale storage systems. Deduplication eliminates redundant data, improves storage utilization and reduces storage cost. This dissertation addresses the challenges of modern data management, primarily caused by the abundance of digital data and the issues of data redundancy. The primary goal is to optimize data management and storage by methodically identifying and eliminating duplicate files using the CRC32b secure hashing technique. This technique generates unique hash values for each data piece, enabling precise identification and removal of duplicate files without compromising data security or integrity. The study evaluates the suitability of the CRC32b secure hashing method for data de-duplication across different use cases. It includes an in-depth analysis of the CRC32b algorithm and the development of a data de-duplication system capable of managing diverse scenarios, such as personal data management, cloud storage, and enterprise-level data centres. The system’s performance is assessed based on quantitative metrics such as processing time, resource usage, and storage space savings. This research provides valuable insights into data redundancy, storage inefficiency, data integrity, and security, offering a comprehensive understanding of these complex issues. The use of the CRC32b secure hashing technique for data de-duplication is innovative and offers a new perspective on tackling data redundancy. This broad scope ensures that the findings are widely applicable and beneficial to a range of data management contexts.

1. Aim –

The aim of this study is to analyze and categorize deduplication techniques in large storage systems, explore their application in data sharing while maintaining privacy, and highlight future research directions in this field.

1. Research Questions –
   1. How does deduplication 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 deduplication techniques categorized based on storage, application point, and level?
   4. What are the potential future research directions in the field of deduplication for academia and industry?
2. Objective –
   1. The role of deduplication technique to improve the performance of a large storage system has been discussed. The necessity of a deduplication technique, its merits and demerits has also been studied.
   2. This scheme can flexibly support data sharing with reduplication even when the data holder is offline, and it does not intrude the privacy of data holders
   3. The existing deduplication techniques have been categorized as storage based, point of application based and level based
   4. Future research directions in the field of deduplication have been highlighted for researchers of academia and industry.