1. **Introduction**
2. **Literature Review**
   1. **Introduction to Data De-Duplication**

As data volumes grow exponentially in cloud storage services, duplication becomes a significant issue. Duplication is used for reliability, availability, and disaster recovery, but excessive duplication can strain the storage system in terms of space and bandwidth. To control this, deduplication techniques are applied, making the storage system more efficient.

Deduplication is a technique that automatically removes duplicate data in storage systems. The application of deduplication techniques varies depending on the type of data, such as structured, unstructured, and semi-structured data, which can further be classified as text, image, and video. Duplicate data can affect storage performance and network bandwidth, leading researchers to focus on developing optimized deduplication techniques. These techniques include the deletion of duplicated data and efficient data delivery in a storage system. Studies have shown that deduplication can achieve significant space savings. For instance, a study conducted on a file system over a period of four weeks found that whole-file-level deduplication achieves 75% of the space savings, while block-level deduplication achieves 32% of the original requirements. Another study applied deduplication on a digital library and found that the quality of metadata deduplication improved significantly. Reports also suggest that deduplication can reduce up to 95% of duplicate data in storage systems, with typical savings being 95% for backup, 72% for VMware, 30% for email, and 35% for file services (Kaur et al.).

With the rapid growth of digital data, particularly from mobile usage and social media, data deduplication has become crucial for eliminating redundant data segments and managing the vast volumes of data. Fingerprints are used in data deduplication to identify identical data blocks. However, as data volume increases, so does the number of fingerprints, leading to a slowdown in data deduplication due to limited memory. Many deduplication solutions face a bottleneck in matching lookups and chunk fingerprint calculations, which require significant storage and processing power. Therefore, using a fast hash algorithm to enhance fingerprint lookup performance is a compelling challenge. This study aims to improve the deduplication system by proposing a novel mathematical bounded linear hashing algorithm. This algorithm reduces the hashing time by over two times compared to MD5 and SHA-1 and decreases the size of the hash index table by 50%. Given the large number of chunk hash values, looking up and comparing hash values becomes time-consuming for large datasets. This work proposes a hierarchical fingerprint lookup strategy to reduce the hash judgement comparison time by up to 78%. The proposed system addresses the high latency caused by deduplication procedures, primarily the hashing and matching phases. The balance of this work lies in the performance of the proposed hashing algorithm and its impact on system efficiency, as well as evaluating the approximate symmetries of the hashing and lookup phases compared to other deduplication systems (Saeed and George).

* 1. **Hashing Techniques in Data De-Duplication**

Data deduplication is a technique that eliminates duplicate data to conserve storage space, and it’s particularly useful in cloud storage. This research focuses on cloud computing and addresses the challenge of removing duplicate files on the cloud while maintaining data security. It’s suggested that data should be encrypted and decrypted to prevent unauthorized access. Existing systems often lack accuracy and security. The proposed approach differs from traditional deduplication systems by giving each client exclusive rights to their data. In the proposed hybrid cloud architecture, the S-CSP is hosted in the public cloud. According to a security study, this research plan is reliable in terms of data protection (Aishwarya et al.).

Biomedical industries generate vast amounts of medical data daily, necessitating substantial storage space and processing power. Cloud computing offers the scalability needed to handle this data. However, securely sharing medical images, a common form of data in these industries, is a significant challenge. The proposed solution is a modified digital signature scheme that ensures the integrity and authenticity of the biomedical image. A third-party auditor checks the shared medical image for damage or modifications using this digital signature scheme. The proposed scheme uses an existing Adler32 hash function to create a digital signature with 8 bits, but it provides a 16-bit output signature, an improvement over existing digital signatures. The proposed scheme has demonstrated its efficiency through various experiments, proving its potential for ensuring the secure sharing of medical images in the cloud. This approach addresses the critical issues of data integrity and authenticity in the sharing of biomedical images, making it a promising solution for secure data sharing in biomedical industries. (Prabhu Kavin et al.).

Persistent memory (PM) is crucial for storing edge data of IoT devices, but its capacity is limited by the embedded system’s host space and energy resources. To enhance PM utilization and energy efficiency, a deduplication system is needed. However, current inline deduplication algorithms cause significant latency when applied to PM file systems (PMFSs) due to low tolerance to the extra deduplication I/O path, lengthy fingerprint calculation, and unnecessary duplication detection for content with low duplication ratio. To address this, LO-Dedup, a novel low-overhead deduplication system for PM, has been proposed. LO-Dedup uses a PM-friendly design that includes fine-grained data load/store and a fast hash scheme to reduce overhead. It can also adaptively sample duplication detection based on recent contents. A prototype of LO-Dedup was implemented in PMFS, a well-known PMFS. Experimental results showed that the write performance only experienced a slight drop when writing data and saved up to 45% space in PMFS. This suggests that LO-Dedup could be a promising solution for improving the efficiency of PM in IoT devices (Chen et al.).

* 1. **CRC32b Hashing Technique**

Time synchronization is crucial for the control and automation of distribution networks. The Precision Time Protocol (PTP), a high-precision protocol based on Ethernet, is commonly used for this purpose. However, it faces challenges such as time jitter and reliability of timestamp verification. A proposed solution is a hashing and verification method that operates at the lower network layers. This method uses a random function with additional interpolation to significantly reduce the collision rate. Experimental results show a collision rate of 0.015%, which can be further reduced to 0.012% with additional interpolation. This is a significant improvement over the Cyclic Redundancy Check (CRC) method used in PTP, reducing the collision rate by 75% to 80%. The proposed method also improves efficiency by about four times compared to CRC running in the Transmission Control Protocol (TCP) stack and system kernel, by bypassing the system stack (Sicheng and Cui).

P4 is a standardized method for programming data planes. However, current P4 targets lack support for processing payload data and specialized cryptographic hash functions, which are essential for secure and resilient communication. This limitation prevents the implementation of applications and protocols requiring message authentication codes or hashing structures resilient against attacks like denial-of-service. The text proposes extending P4 targets with cryptographic hash functions to enable authentication and resilience. An extension of the P4 Portable Switch Architecture for cryptographic hashes is suggested, with prototype implementations for three different P4 target platforms: CPU, NPU, and FPGA. Performance evaluation and resource consumption analysis of these prototypes show that cryptographic hashing can be efficiently integrated. However, no single hash function delivers satisfactory performance on all platforms, leading to the recommendation of a set of hash functions for optimizing target-specific performance (Scholz et al.).

* 1. **Security in Data De-Duplication**

Cloud service providers aim to offer efficient and effective storage and data transmission. To save on storage costs and bandwidth, they use data de-duplication. However, users often encrypt their data for privacy and security before uploading it to the cloud, which conflicts with the de-duplication function. Existing de-duplication methods face challenges in terms of security and efficiency, being either vulnerable to brute force attacks or computationally expensive. This text proposes a performance and security-effective method for data de-duplication. It begins with an overview of de-duplication strategies, then discusses the security and efficiency issues of existing approaches. An enhancement is proposed using the AES-CBC algorithm and hashing functions to improve the performance and security of data de-duplication. This method allows for the safe and consistent creation of users' keys without third-party involvement. The effectiveness of the proposed solution is demonstrated through practical application and comparison with existing techniques. The keywords associated with this text are Data duplication, De-duplication, Cloud computing, Security, and Encryption (Et al.)

Cloud computing includes a service known as cloud storage, where data is remotely maintained, managed, and backed up. With the exponential increase in data, issues related to storage space, data confidentiality, and search space complexities also increase. The proposed model aims to address these issues by storing data redundantly using an efficient de-duplication technique and protecting the confidentiality of sensitive data. It also monitors activities in the storage environment to provide security to the storage nodes. As these nodes are geographically distributed, the focus is on optimal data storage and retrieval, storage management, and data security. The model improves efficiency in data storage and retrieval, optimal storage allocation, and de-duplication, while also enhancing data security. Given the large volume of data, the efficient storage and retrieval of data in an optimal way is addressed. This optimal storage and retrieval reduce latency, thereby increasing throughput. As the data is stored in a remote location, security is also efficiently addressed (Ashmita and Anitha).

* 1. **Challenges in Data De-Duplication**

Content-Defined Chunking (CDC) is crucial in data deduplication systems due to its high redundancy detection ability. However, existing CDC-based approaches have high CPU overhead as they compute and judge the rolling hashes of the data stream byte by byte. FastCDC, a fast and efficient CDC approach, is proposed for data deduplication-based storage systems. FastCDC uses five key techniques: gear-based fast rolling hash, simplified and enhanced Gear hash judgment, skipping sub-minimum chunk cut-points, normalizing the chunk-size distribution in a small specified region, and rolling two bytes each time. These techniques make FastCDC 3-12X faster than other CDC approaches, while maintaining or even improving the deduplication ratio. FastCDC-based Destor, an open-source deduplication project, achieves 1.2-3.0X higher throughput than Destor based on other chunkers (Xia et al.).

* 1. **Related Work in Data De-Duplication**

Data de-duplication is a method that allows cloud providers to manage the vast amounts of data generated by the Internet of Things (IoT) and social media, which present significant challenges for cloud storage. While big data offers business opportunities, it also poses substantial issues for cloud service providers, particularly in terms of data storage and management. This discussion focuses on the problem of redundant data and explores techniques to prevent data redundancy in the cloud. A version control system for data de-duplication is also proposed as a solution to this issue (Xia et al.).

In the context of data de-duplication, a system is designed to back up data. This system utilizes client software installed on the user’s computer. The software identifies all files necessary for backup and transforms them into a byte stream. This stream is then segmented into blocks, each ranging from 32 KB to 64KB, using a Rabin algorithm. This algorithm, based on a hash ring, processes each incoming byte. When the current hash mask matches a certain threshold or reaches 64KB, a division is performed. This method helps to prevent the need for duplicating all data if a file’s content has been modified independently. For each data block, the client software computes the hash. It then sends portions of these hashes, potentially up to 256 times or more, to the server to verify if they are already recognized by the system. Blocks that are not recognized are sent to the server. During the de-duplication process, data is distributed on the cloud based on these hashes. These hashes, generated using SHA-1, produce 20 bytes in hexadecimal format. The first character of the hash is used as the key for distribution, allowing for an even distribution of data among the workers, ranging from 2 to 16 pieces. For additional distribution, the second character of the hash is utilized. This approach ensures a balanced and efficient system for data de-duplication (Lytvyn).

* 1. **Future Scope**

One of the primary areas of focus could be the enhancement of security measures. As data security becomes a paramount concern in today’s digital age, there could be extensive research into developing more secure hashing techniques. This could involve the creation of innovative algorithms or the enhancement of existing ones like CRC32b. Another potential area for future work could be the optimization of performance. This would involve improving the speed and efficiency of the hashing and de-duplication processes, thereby making the system more robust and reliable. Scalability is another crucial aspect that could be explored in future research. With the exponential growth in data volume, there is a pressing need for data de-duplication systems that can scale efficiently. This would involve devising techniques that can handle larger datasets without compromising on performance. The demand for real-time data de-duplication solutions is also likely to grow in the future. This would necessitate the development of systems that can perform de-duplication quickly and efficiently as data is being generated or processed. Integration with other technologies could also be a significant area of future work. For instance, data de-duplication systems could be combined with machine learning algorithms to improve the accuracy of de-duplication, opening up new avenues for research. Finally, the development of industry-specific applications could also be a potential area of focus. For instance, in the healthcare industry, efficient de-duplication of patient records could be a key area of focus. These are just a few possibilities. The field of data de-duplication is dynamic and constantly evolving, offering numerous opportunities for future research and development.

**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.