Universiti Teknologi MARA

Data Deduplication Using Hashing Algorithm

Naimah binti Nayan

Thesis submitted in fulfilment of the requirements for

Bachelor of Computer Science (Hons) Data Communication and Networking

Faculty of Computer and Mathematical Sciences

DECEMBER 2018

STUDENT DECLARATION

I certify that this thesis and the project to which it refers is the product of my own work and that any idea or the quotation from the work of the other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline.

……
NAIMAH NAYAN
2016706961

DECEMBER 4, 2018

ABSTRACT

Data deduplication is method that help reduce the redundant data in storage capacity. With the rapid growth of digital data that generated in the digital world, the capacity of storage usage will increase rapidly. To achieve deduplication efficiency in system storage, the duplicate data need to be eliminated. To eliminated the duplicate data, the file unique value or hash value need to compare and the files that have the same hash value will be remove. This method basically will help to improve the storage capacity and efficiency. The hash value is generated by using hashing algorithm such as Message Digest 5 (MD5) and Secure Hashing Algorithm 1 (SHA-1). The hash functions should not create the same index value for the different data. If there is lack of analysis on the hashing algorithm, the deduplication technique cannot be improved for future research and the evolution of data deduplication can be slow because the performance metric for each hashing algorithm is not clear enough. The objective of this project is to compare MD5 & SHA-1 algorithm in data deduplication technique and to evaluate the MD5 & SHA-1 algorithm, length of message digest and speed using deduplication software. The simulation was conducted using File Alyzer, Clone Files Checker and AllDup software. The result of this simulation had been analysed based on three performance metrics which is efficiency, message digest length and the speed. There were two type of dataset which is video and document files with four different sizes. The time taken of the hashing algorithm generate the hash value were recorded. The findings in this project is the MD5 speed performance is better than SHA-1 hashing algorithm because it generates the hash value faster due to the length of message digest in MD5 is shorter than SHA-1. The recommendation for future work is to evaluate various type of data and different type of hashing algorithm.

TABLE OF CONTENTS