

**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 a method that helps reduce the redundant data in storage capacity. With the rapid growth of digital data that is 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 eliminate the duplicate data, the file unique value or hash value need to be compared and the files that have the same hash value will be removed. This method basically will help to improve the storage capacity and efficiency. The hash value is generated by using hashing algorithms 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 a 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 algorithms in data deduplication techniques and to evaluate the MD5 & SHA-1 algorithms, length of message digest and speed using deduplication software. The simulation was conducted using File Alyzer, Clone Files Checker and AllDup software. The results of this simulation have been analysed based on three performance metrics which are efficiency, message digest length and speed. There were two types of datasets which are video and document files with four different sizes. The time taken for the hashing algorithm to generate the hash value was recorded. The findings in this project are that the MD5 speed performance is better than the SHA-1 hashing algorithm because it generates the hash value faster due to the length of the message digest in MD5 being shorter than in SHA-1. The recommendation for future work is to evaluate various types of data and different types of hashing algorithms.

# TABLE OF CONTENTS

CONTENTS	PAGE
SUPERVISOR APPROVAL	i
STUDENT DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	vii
LIST OF TABLES	viii
CHAPTER 1	1
1.1 Background of Research Study	1
1.2 Problem Statement	2
1.3 Research Objectives	4
1.4 Research Scope	4
1.5 Research Significance	4
CHAPTER 2	5
2.1 Data Deduplication	5
2.2 Deduplication Techniques Classification	7
2.3 Deduplication Approaches	8
2.4 Chunking	9
2.4.1 File-Level Chunking	10
2.4.2 Block Level Chunking	10
2.4.3 MD5 Algorithm	11
2.4.4 SHA-1 Algorithm	12
2.5 Related Work	12
2.5.1 A Comparative Analysis of SHA and MD5 Algorithm	12
2.5.2 Bimodal Content Defined Chunking for Backup Streams	13
2.5.3 Improving Accessing Efficiency of Cloud Storage Using Deduplication and Feedback Schemes	13
2.5.4 Improving Restore Speed for Backup Systems That Use Inline Chunk Based Deduplication	13
2.5.5 Evaluation of Two Thresholds Two Divisor Chunking Algorithm Using Rabin Finger print, Adler, and SHA1 Hashing Algorithms	14
CHAPTER 3	15
3.1 Initiation Phase	15
3.2 Project Requirement Phase	16
3.1.1 Hardware Requirement	17
3.1.2 Software Requirement	17
3.3 Design Phase	18
3.4 Analysis Phase	19
3.5 Documentation Phase	19
CHAPTER 4	21
4.1 File Alyzer Installation	21
4.2 Clone Files Checker Installation	23
4.3 AllDup Installation	26

<b>4.4</b>	<b>Performance Measure</b>	<b>28</b>
<b>4.4.1</b>	<b>Efficiency of The MD5 &amp; SHA-1</b>	<b>29</b>
<b>4.4.2</b>	<b>Length of Message Digest of The MD5 &amp; SHA-1</b>	<b>29</b>
<b>4.4.3</b>	<b>Speed of The MD5 &amp; SHA-1</b>	<b>29</b>
<b>4.5</b>	<b>Summary</b>	<b>29</b>
	<b>CHAPTER 5</b>	<b>30</b>
<b>5.1</b>	<b>Analysis of MD5 and SHA-1 Hashing Algorithm.</b>	<b>30</b>
<b>5.2</b>	<b>Scenario1: Analysis of MD5 based on video files dataset.</b>	<b>30</b>
<b>5.3</b>	<b>Scenario2: Analysis of SHA-1 based on video files dataset.</b>	<b>31</b>
<b>5.4</b>	<b>Scenario3: Analysis of MD5 based on document files dataset.</b>	<b>32</b>
<b>5.5</b>	<b>Scenario4: Analysis of SHA-1 based on document files dataset.</b>	<b>33</b>
<b>5.6</b>	<b>Comparison of speed between MD5 and SHA-1 based on video files dataset.</b>	<b>34</b>
<b>5.7</b>	<b>Comparison of speed between MD5 and SHA-1 based on document files dataset.</b>	<b>35</b>
<b>5.8</b>	<b>Comparison of message digest length between MD5 and SHA-1 hashing algorithm.</b>	<b>36</b>
<b>5.9</b>	<b>Summary</b>	<b>37</b>
	<b>CHAPTER 6</b>	<b>38</b>
<b>6.1</b>	<b>Conclusions</b>	<b>38</b>
<b>6.2</b>	<b>Recommendation for future work</b>	<b>38</b>
	<b>REFERENCES</b>	<b>40</b>