MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

DATA DEDUPLICATION USING VARIABLE LENGTH CHUNKING

B.Tech. Major Project Report

SUBMITTED BY

Shraddha Jadhav	**[B194065]**
Jagath Sai Narayana	**[B194075]**
Shardul Dubey	**[B194100]**
Shubham Suryawanshi	**[B194087]**

GUIDED BY

Prof. Amar More

SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY

MIT ACADEMY OF ENGINEERING, ALANDI (D), PUNE-412105

MAHARASHTRA (INDIA)

MAY, 2020

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

DATA DEDUPLICATION USING VARIABLE LENGTH CHUNKING

A Major Project Report

submitted in fulfilment of the
requirements for the award of the degree

of

Bachelor of Technology

in

COMPUTER ENGINEERING

SCHOOL OF COMPUTER ENGINEERING & TECHNOLOGY

MIT ACADEMY OF ENGINEERING, ALANDI (D), PUNE-412105

MAHARASHTRA (INDIA)

MAY, 2020

MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

CERTIFICATE

It is hereby certified that the work which is being presented in the B.Tech. Major Project Report entitled **“***Data Deduplication using Variable Length Chunking***”,** in fulfillment of the requirements for the award of the **Bachelor of Technology in Computer Engineering & Technology** and submitted to the **School of Computer Engineering & Technology of MIT Academy of Engineering, Alandi(D), Pune, Affiliated to Savitribai Phule Pune University (SPPU), Pune** is an authentic record of work carried out during an Academic Year 2019-2020, under the supervision of **Prof.Amar More School of Computer Engineering & Technology**.

Shraddha Jadhav	**PRN: 0120160369**	**Seat No.: B194065**
Jagath Sai Narayana	**PRN: 0120160422**	**Seat No.: B194075**
Shardul Dubey	**PRN: 0120160520**	**Seat No.: B194100**
Shubham Suryawanshi	**PRN: 0120160461**	**Seat No.: B194087**

Date:

Signature of Project Advisor
Project Adviser
School of Computer Engineering & Technology
MIT Academy of Engineering, Alandi(D), Pune

Signature of Dean
Dean
School of Computer Engineering & Technology
MIT Academy of Engineering, Alandi(D), Pune

(STAMP/SEAL)

Signature of Internal examiner/s
Name……………………………
Affiliation…………………………

Signature of External examiner/s
Name……………………………
Affiliation…………………………

ACKNOWLEDGEMENT

Every orientation work has an imprint of many people and it becomes our duty to express deep gratitude for the same.

During the entire duration of this seminar, we received endless help from a number of people and we feel that this report would be incomplete if we don't convey thanks to them. This acknowledgement is a humble attempt to thank all those who were involved in the project work and were of immense help to us.

We want to express our gratitude towards our respected project guide Prof.Amar More for his constant encouragement and valuable guidance during the completion of this project work. We also want to express our gratitude towards respected School Dean Mrs. Ranjana Badre for her continuous encouragement.

We would be failing in our duty if we do not thank all the other staff and faculty members for their experienced advice and evergreen co-operation.

1. Shraddha Jadhav sign
2. Jagath Sai Narayana sign
3. Shardul Dubey sign
4. Shubham Suryawanshi sign

DECLARATION

We the undersigned solemnly declare that the project report is based on our own work carried out during the course of our study under the supervision of Prof Amar More.

We assert the statements made and conclusions drawn are an outcome of our research work. We further certify that

1) The work contained in the report is original and has been done by us under the general supervision of our supervisor.
2) The work has not been submitted to any other Institution for any other degree/diploma/certificate in this Institute/University or any other Institute/University of India or abroad.
3) We have followed the guidelines provided by the Institute in writing the report.
4) Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and giving their details in the references.

Exam Seat No.	**Name**	**Sign**
B194065	Shraddha Jadhav	
B194075	Jagath Sai Narayana	
B194100	Shardul Dubey	
B194087	Shubham Suryawanshi	

ABSTRACT

Flexibility and cost efficiency provided by cloud storage vendors like Amazon, Google Cloud Platform, etc. are attracting many organizations for migrating their data to the cloud storage . Also the number of people using social media like Facebook, Twitter, WhatsApp have already crossed the count of some billions. The amount of the data posted by the people on those social media is also increasing exponentially. The studies have shown that, among the data posted by people, more than 50 percent of the data is duplicate. If we think from cloud storage provider point of view, then storing duplicate data require more storage space and energy which actually is a waste. If we could detect this duplicate data and store only one copy of it, then lot of space and energy will be saved. For maintaining the reliability of the data, the storage providers will have to replicate data, thereby generating the duplicate data. Thus reliability and deduplication are two sides of one coin and if handled efficiently, will help in reducing the extra space and energy to store data and also provide the reliability. In this project, we propose energy efficient reliability aware distributed data deduplication for storing data. The algorithm will detect the duplicate data from the data stored on many servers, and will maintain only one copy of the data and to provide reliability, the algorithm will maintain the multiple copies of this to achieve both deduplication and reliability. We make use of content defined chunking to detect the duplicate data more efficiently and use distributed hash tables to reduce the read and write latency