11 January 2023

Computer Science, BINUS International University

ADVISOR

Dr. MARIA SERAPHINA ASTRIANI, S.Kom., M.T.I.

STUDENT

ALYSHA P. MAULIDINA
CHELLSHE LOVE SIMROCHELLE
KIMBERLY MAZEL
RACHEL ANASTASIA WIJAYA

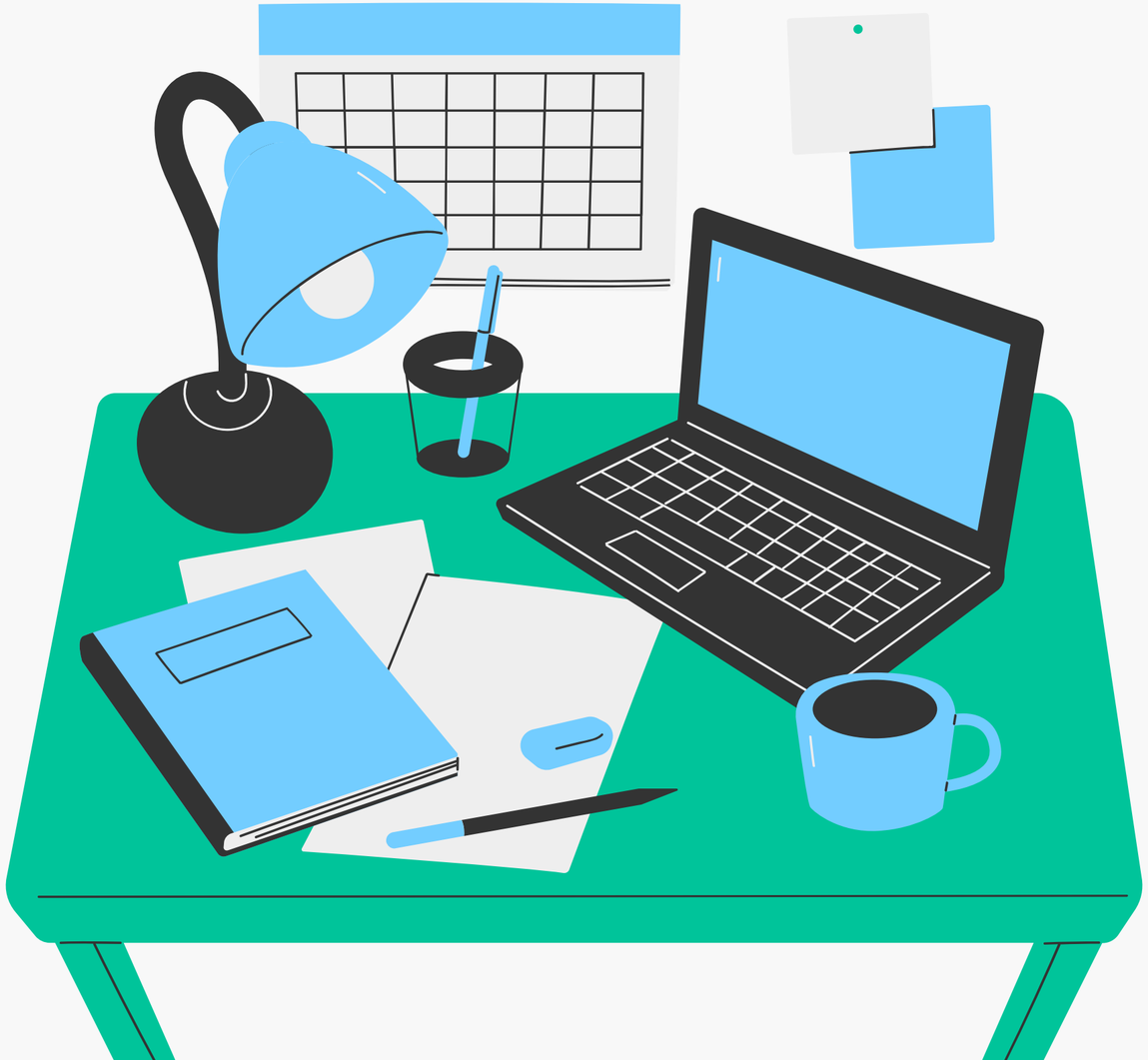
Comparative Study of Data Compression Algorithms

I Introduction: What is data compression?

The conversion of data into a more compact form of representation that has fewer storage requirements.

Why is data compression important?

- Era of rapid communication & transfer of data, Big Data revolution
- Enables the ease of data transmission in the cloud



Many different algorithms available with its own strengths and weaknesses - comparative study can provide insight into the performance of different algorithms and identify the most fastest option for different applications.

Z-standard

"Fast lossless compression algorithm, targeting real-time compression scenarios at zlib-level and better compression ratios"

Uses Huffman coding & Finite State Entropy

zlib

"zlib is a commonly used library built on an old codebase"

Uses DEFLATE, which is based on Huffman coding

LZ4

LZ4 is lossless compression algorithm, providing compression speed "500 MB/s per core"

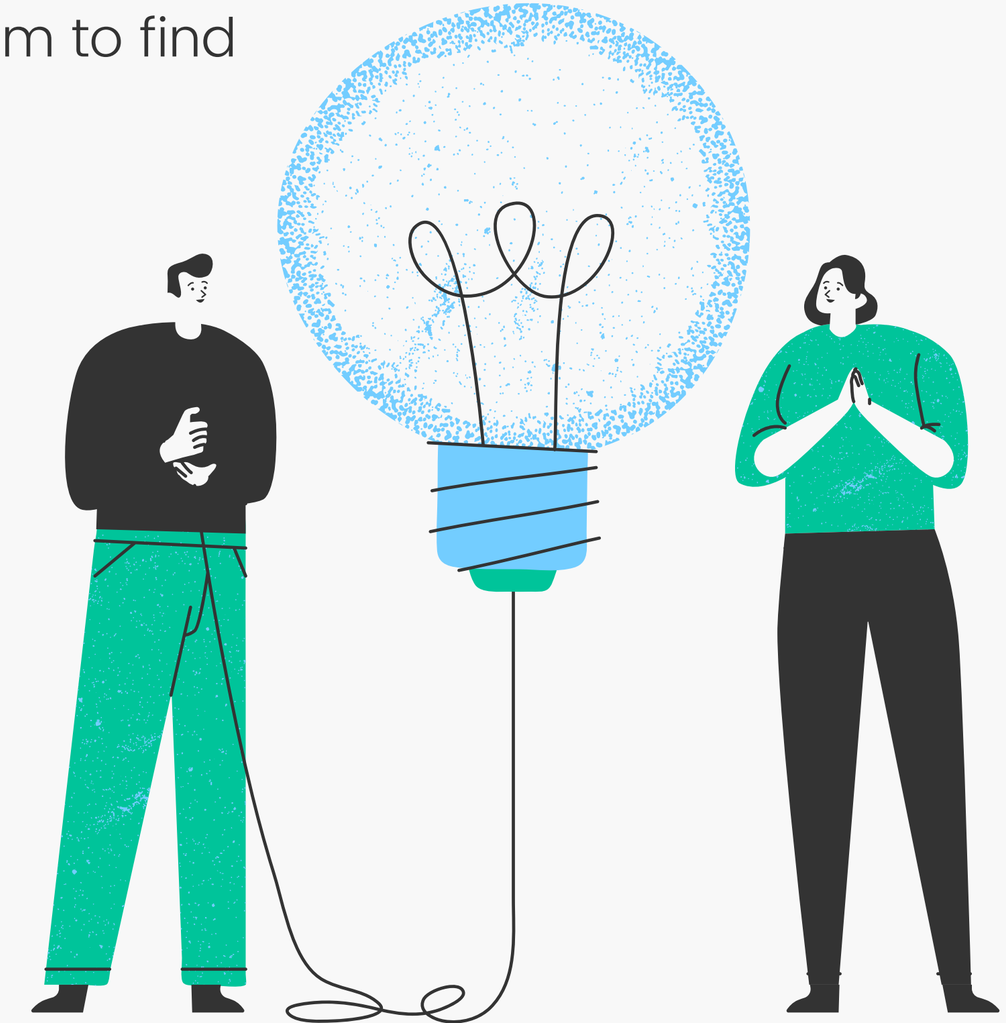
Uses Dictionary compression

III Proposed solution

Testing different algorithms with the same corpus and measuring the performance of each algorithm to find their strengths and weaknesses.

Three algorithms were chosen as they all promise high compression speed & ratios.

We aim to study the process behind each algorithm to further understand the results in depth.



IV Methodology

Step 1

Curate a corpus (test data)

Step 2

Test the algorithm

Step 3

Repeat tests and get the average value

Step 4

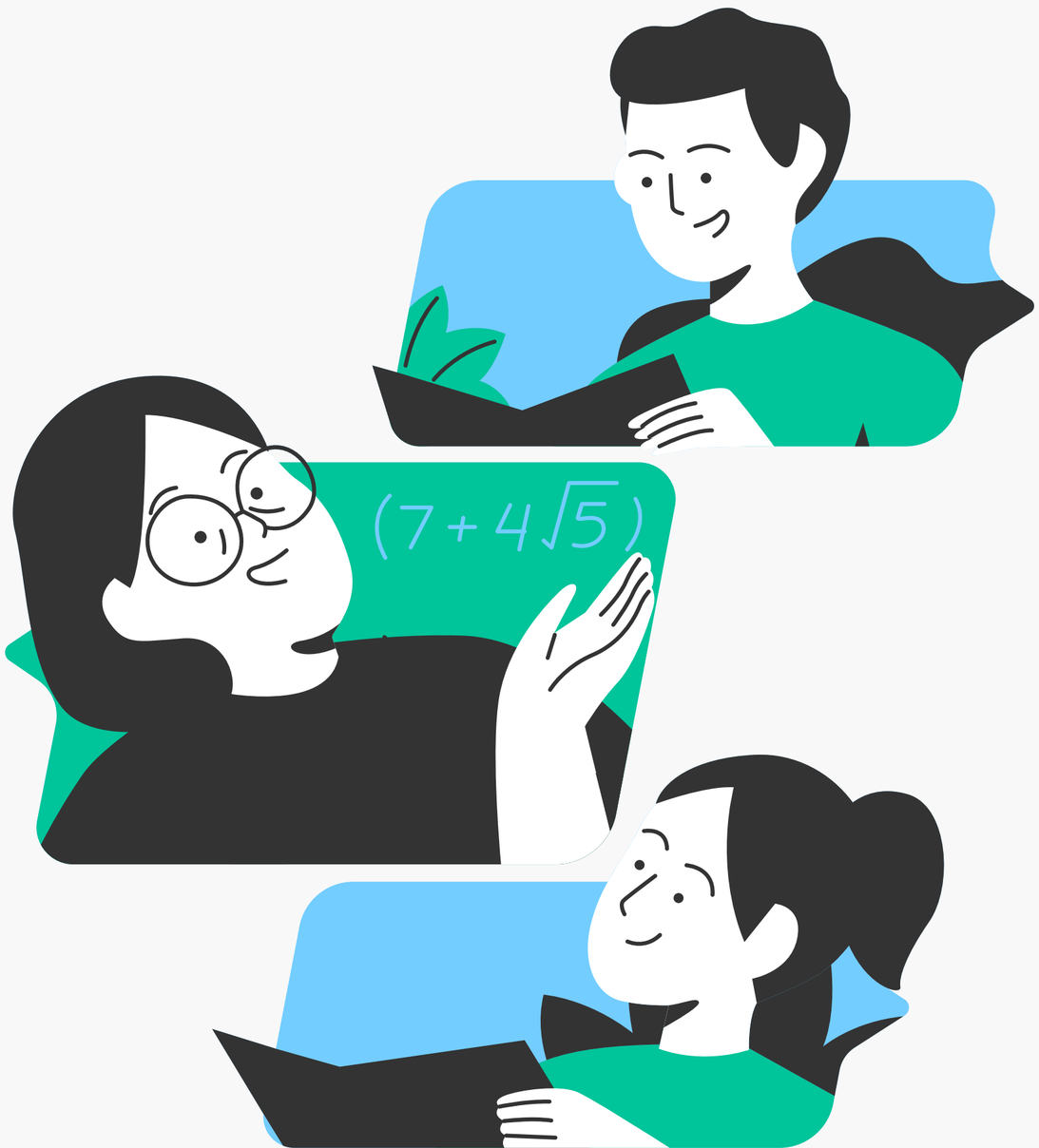
Analyze and conclude the strengths and weaknesses

In order to properly evaluate the different algorithms, certain variables have to be controlled to ensure that the measured variables are only affected by the algorithms alone and not external factors. For example, the test data used for each algorithm is the same, as well as the environment that it is tested on (Computer Specifications)

V Findings

Link to CSV

File Type	Best Algorithm
Text	Zstandard
Small Image	Zlibrary
Big Image	Zlibrary
PDF	Zstandard
CSV	Zstandard
JSON	Zlibrary



Zstandard

Highly fast decoding, consistent compression, it has a bigger overhead and relies on FSE, which is recent and still in development

Zlibrary

Works reasonably well. It is generally slowest in compression speed and is occasionally better in compression ratio (for images) and decompression speed.

LZ4

Worst compression ratio out of the three but have the highest speeds as a result of its small overhead

V Conclusion & Discussions

- Zlib uses the DEFLATE algorithm, a universally adopted lossless compression algorithm used today (also implemented for zip).
- Our findings demonstrate why this algorithm is suitable for general use cases.

- ZStandard is advanced but it is still a newly developed algorithm that is at its infancy stage. The technology may not be as reliable as Zlib as it is not as open.

- LZ4 has the worst compression, but it has a fast decoding speed. It has a small overhead and its decoder is simple, so it is one of the only viable options when dealing with small microcontrollers/microprocessors

11 January 2023

Computer Science, BINUS International University

ADVISOR

Dr. MARIA SERAPHINA ASTRIANI, S.Kom., M.T.I.

STUDENT

ALYSHA P. MAULIDINA
CHELLSHE LOVE SIMROCHELLE
KIMBERLY MAZEL
RACHEL ANASTASIA WIJAYA

Thank you for listening!