**MINOR-1 PROJECT**

**SYNOPSIS on**

**Comparing lossless Data Compression Algorithms**

Submitted By :

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **Branch** |
| Ashi Agarwal | R110216043 | CSE CCVT |
| Arpit Bhardwaj | R110216039 | CSE CCVT |
| Deepanshu Goyal | R110216057 | CSE CCVT |
| Ashish Bansal | R110216044 | CSE CCVT |

**Under the guidance of**

**Mr. G.L. Prakash**

**Asst. Professor (S.G.)**

**Department of Virtualization | SoCS|UPES**



School of Computer Science

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**Dehradun-248007 2018-19**

**Approved By**

**Mr. G.L. Prakash**

**Department of Virtualization**  ()

**Project Guide Department Head**



**School of Computer Science**

University of Petroleum & Energy Studies, Dehradun

**Synopsis Report (2018-19)**

**1 Project Title**

# Abstract

In today’s world, With the advent of the Internet and mobile devices with limited resources and with the growing requirements of information storage and data transfer, Cloud Computing has become an important aspect but cloud computing also require physical infrastructure, somewhere down the lane. This exponential sub purge of data leads to high demand for data processing that leads to a high computational requirement which is usually not available at the user's end. Compression reduces the redundancy in data representation thus increasing effective data density. [1] Data Compression is a technique which is used to decrease the size of data. This is very useful when some huge files have to be transferred over networks or being stored on a data storage device and the size is more than the capacity of the data storage or would consume so much bandwidth for transmission in a network. With the limited physical infrastructure for storage, data compression has gained even more importance these days. There are number of data compression algorithms, which are dedicated to compressing different data formats. Even for a single data type, there are number of different compression algorithms, which use different approaches. In this project, we will examine lossless data compression algorithms like Huffman encoding algorithm, Lempel-Ziv-Welch algorithm, and Shannon-Fano algorithm and comparing their performance.[2]

**Keywords: Cloud Computing, Data Compression ,Huffman encoding algorithm, Lempel-Ziv-Welch algorithm, Shannon-Fano algorithm.**

# Introduction

Compression is the art of representing the information in a compact form rather than its original or in uncompressed form [3]. In other words, using the data compression, the size of a particular file can be reduced.This is very useful when processing, storing or transferring a huge file, which needs lots of resources. If the algorithms used to encrypt works properly, there should be a significant difference between the original file and

the compressed file. When data compression is used in a data transmission application, speed is the primary goal. Speed of transmission depends upon the number of bits sent, the time required for the encoder to generate the coded message and the time required for the decoder to recover the original message. In a data storage application, the degree of compression is the primary concern. There are two types of data compressions ie. Lossless data compression and Lossy data compression.

|  |  |
| --- | --- |
| **Lossless data compression** | **Lossy data compression** |
| In Lossless data compression algorithms the original data can be recovered from compressed data after applyind decompression algorithm | In Lossy compression algorithms it permanently reduces the original data by eliminating certain information, especially redundant information, after decompressing the file only the part of original data is recovered. |
| Lossless compression is generally used for text data or spreadsheet files ,where even a very small amount of data loss can be detected by users. | Lossy compression is generally used for video and sound, where a certain amount of information loss will not be detected by most users. |
| No loss in information so compression rate is small. | In return for accepting this distortion in reconstructed data we obtain high compression rate. |
| Less data can be accommodated in channel. | More data can be accommodated in channel. |
| *E*.*g*(*i*)*F**a**x**M**a**c**h**i**n**e*,(*i**i*)*R**a**d**i**o**l**o**g**i**c**a**l**I**m**a**g**i**n**g* | *E*.*g*.(*i*)*T**e**l**e**p**h**o**n**e**S**y**s**t**e**m*,(*i**i*)*V**i**d**e**o**C**D* |

Various lossless data compression algorithms have been proposed and used. Some of the main techniques in use are the Huffman Coding, Run Length Encoding, Arithmetic Encoding and Dictionary Based Encoding [4]. We will examine the performance of the Huffman Encoding Algorithm,Shannon Fano Algorithm, and Lempel Zev Welch Algorithm. In particular, performance of these algorithms in compressing text data will be evaluated and compared.

**Shannon-Fano Algorithm:**

In Shannon–Fano Algorithm, the symbols are arranged in order from most probable to least

probable, and then divided into two sets whose total probabilities are as close as possible to

being equal. All symbols then have the first digits of their codes assigned; symbols in the first set receive "0" and symbols in the second set receive "1". As long as any sets with more than one member remain, the same process is repeated on those sets, to determine successive digits of their codes. When a set has been reduced to one symbol, of course, this means the symbol's code is complete and will not form the prefix of any other symbol's code.The algorithm works, and it produces fairly efficient variable-length encodings; when the two smaller sets produced by a partitioning are in fact of equal probability, the one bit of information used to distinguish them is used most efficiently. Unfortunately, Shannon–Fano does not always produce optimal prefix codes.[5]

**Huffman Encoding:**

The Huffman encoding works on variable length encoding rather than fixd length encoding. frequency of every character is calculated and the lengths of the assigned codes are based on the frequencies of corresponding characters. The most frequent character gets the smallest code and the least frequent character gets the largest code.. .The variable-length codes assigned to input characters are Prefix Codes, means the codes (bit sequences) are assigned in such a way that the code assigned to one character is not prefix of code assigned to any other character. This is how Huffman Coding makes sure that there is no ambiguity when decoding the generated bit stream For this task a binary tree is created using the symbols as leaves according to their probabilities and paths of those are taken as the code words

**The Lempel Ziv Welch Algorithm:**

It is a dictionary based compression algorithm As in Dictionary the set of all possible word of a language is stored similarly in lzw algorithm a dictionary is used to store or index the previously used string patterns In the compression process those index value is used instead of repeating . The dictionary is created dynamically in the compression process and no need to transfer it with the encoded message for decompressing. In the decompression process, the same dictionary is created dynamically. Therefore, this algorithm is an adaptive compression algorithm.

**Performance of compression algorithm** [6] is based on space efficiency and the time complexity

The compression behavior of algorithm is dependent on redundancy of symbol in source file therefore it is difficult to measure the performance of compression algorithm. There are some following measurements used to evaluate the performances of compression algorithms.

**Compression Ratio** –Compression Ratio is the ratio between the size of the compressed file and the size of the source file

Compression ratio= (size of compressed file )/(size of source file)

**Compression Factor**-Compression factor is inverse of Compression ratio It tales how much time our file has been compressed

Compression Factor= (size of source file)/(size of compressed file)

**Saving Percentage**-It tales the shrinkage of the source file in percentage

Saving Percentage= (size before compression-size after compression)/(size before compression)%

**Time complexity** –The time complexity is measured by the number of clock used to encode or decode the source code .The algorithm that uses less clock cycle to encode or decode is considered more efficient.

**Code Efficiency** Average code length is the average number of bits required to represent a single code word. If the source and the

lengths of the code words are known, the average code length can be calculated using the following equation.



,where pjis the occurrence probability of j th symbol of the source message, l j is the length of the particular code word for that symbol and L = {l 1 , l 2 , .... L n }.

In this project we will be compairing all the three algorithms on the basis of these factors mentioned above.

# Literature Review

# Problem Statement

# Objectives

# Methodology

# System Requirements (Software/Hardware)

# Schedule (PERT Chart)

# References

[1] Monika Soni , Dr Neeraj Shukla “Data Compression Techniques in Cloud Computing”

[2] Mohammad Hosseini “A Survey of Data Compression Algorithms and their Applications”

[3] Pu, I.M., 2006, Fundamental Data Compression, Elsevier, Britain.

[4] Kesheng, W., J. Otoo and S. Arie, 2006. Optimizing bitmap indices with efficient compression, ACM Trans. Database Systems, 31: 1-38.

[5] <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=15&ved=2ahUKEwjW8piM_q_dAhVHyrwKHXE2DHoQFjAOegQIABAC&url=http%3A%2F%2Fecehithaldia.in%2Fteaching_material%2FShanon-Fano1586521731.pdf&usg=AOvVaw0MHM4foSS-sDhzqyRAVfaE>

[6] S.R. Kodituwakku ,U. S. Amarasinghe “Comparision of Lossless data compression algorithms for text data”