File compression

- the process of reducing the size of a file or a collection of files to save storage space

- or decrease the time required to transmit the files over a network

- achieved by using compression algorithms

- compression algorithms: identify and remove redundancy in the data (thus representing the information more efficiently)

- Redundancy Removal: repeated patterns, duplicate sequences, or predictable information

- Entropy Coding: represent data more efficiently

assigning shorter codes to more frequent symbols

longer codes to less frequent symbols

eg. Huffman coding and arithmetic coding

-Dictionary-Based Compression: use dictionaries

replace repeated patterns with references to a dictionary entry

eg. Lempel-Ziv (LZ) variants

- Run-Length Encoding (RLE): replaces sequences of identical symbols with a single symbol followed by a count of the number of repetitions

- Lossless Compression: retains all the original data

ensuring that the decompressed file is an exact replica of the original

eg. ZIP, Gzip, and Bzip2

- Lossy Compression: sacrifices some level of data precision to achieve higher compression ratios

often used in multimedia compression

eg. JPEG for images, MP3 for audio, and MPEG for video

- Compression Tools:

- ZIP: (.zip)

one of the most widely used compression formats

supports lossless data compression and includes both compression and archiving functionalities

ZIP files can be created and extracted using various tools on different operating systems

- Tar: (.tar)

Tar is not a compression tool by itself but is often used in combination with other compression tools

creates uncompressed archives (tarballs)

additional tools like gzip or bzip2 are used to compress the tarball

combined file is typically given a .tar.gz or .tar.bz2 extension.

- Gzip: (.gz)

command-line compression tool often used in Unix and Linux environments

uses the DEFLATE compression algorithm and is commonly used to compress individual files

Files compressed with gzip have the extension .gz.

- Bzip2: (.bz2)

command-line compression tool that uses the Burrows-Wheeler Transform and Run-Length Encoding

generally provides higher compression ratios than gzip but may have slower compression and decompression speeds

Files compressed with bzip2 have the extension .bz2.

- WinZip & WinRAR

- let's compress and extract some file with zip and tar

// install zip and unzip

sudo apt install zip

sudo apt install unzip

zip archive_file_name.zip [absolute/relative path to the file or directory]

eg. zip day1.zip ./day1

it only compress the specified folder but not the contents within

to compress whole folder, we need recursive function (-r) implemented

eg. zip -r Day2b.zip ./Day2b

'-j' is used after zip name if just files are to be compressed not the folder structure

multiple folder's file into the single zip

eg. zip -r mul.zip -j ./day1 ./Day2b ./day5

unzip archive_file_name.zip -d [absolute/relative path to the destination]

eg. unzip day1.zip -d ./day2b

// install tar

sudo apt install tar

tar [options] file_name.tar [absolute/relative path to the file or directory]

options: -c = create, -x = exract, -v = show progress, -f = specify filename ourself, -z = .gz, -j = .bzip2, [--exclude='filename.ext'] : to exclude certain file

tar -cf arc.tar ./day1

tar -xf arc.tar ./Day2b

Automation of script and scheduling the script using cron:

- Cron is a time-based job scheduler in Unix-like operating systems
- allows users to schedule tasks (commands or scripts) to run periodically at predefined intervals
- Guide: Search for 'guide on how to automate tasks with cron' in chatGPT
- Access the Cron Table:

crontab -e

// result

no crontab for rrp - using an empty one

Select an editor. To change later, run 'select-editor'.

1. /bin/nano <---- easiest
2. /usr/bin/vim.basic
3. /usr/bin/vim.tiny
4. /bin/ed

Choose 1-4 [1]:

- Cron Syntax (inside corn table):

minute hour day month day_of_week [Script Name]

minute: 0 to 59

hour: 0 to 23

day: 1 to 31

month: 1 to 12

day_of_week: 0 to 6 (Sunday to Saturday)

- common symbols:

*: Represents all possible values for a field.

*/n: Represents every nth value.

,: Represents a list of values.

-: Represents a range of values.

- Viewing the Current Cron Jobs:

crontab -l