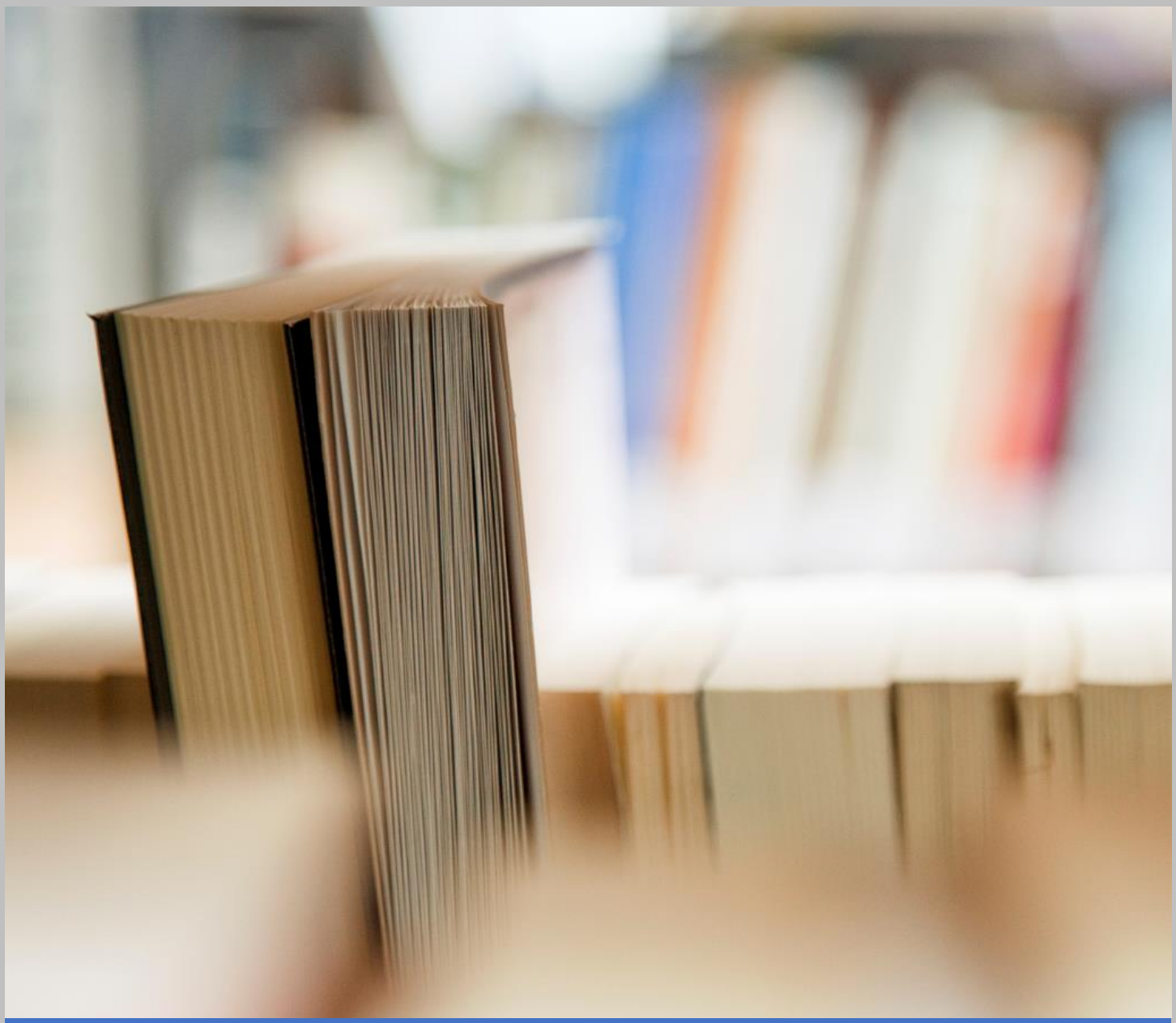


File archiving and compression using tar, gzip and biz/b2z

ankit shivane

7/24/24

DevOps Connect

Tar utility

Archiving and compressing files are common operations in the Unix world, done by system administrators on a very regular basis.

Luckily for you, Linux exposes a set of different commands in order to archive, compress, uncompress and extract files from an archive.

A. Archive files on Linux using tar

Tar is a very popular command among system administrators.the tar command does not compress .tar files. tar just bundles up

the files into a single archived file. Therefore, if you ever see a file ending in just '.tar', you already know that

the archive process applied no compression on the files contained within that archive.

The tar command has a few more options than the zip command did. The most commonly used options for the tar command

includes the following:

-c: creates a new .tar archive file

-v: verbosely shows the tar process so you can see all the steps in the process

-f: specifies the file name type of the archive file

-x: extracts files from an existing .tar file

-t: --list, it will list the contents of an archive

-u: append new file to the existing archive file that is .tar file

a.

The following example shows the basic syntax of the tar command to create an archive:

$ tar -cvf <name of archive file>.tar <directory to archive or files to archive>

tar -cvf archive.tar file1 file2 file3

b. list the content of archive file using -t:

[ec2-user@ip-172-31-15-192 p2]$ tar -tvf archive.tar

-rw-r--r-- ec2-user/ec2-user 19559749 2024-07-24 09:53 file3

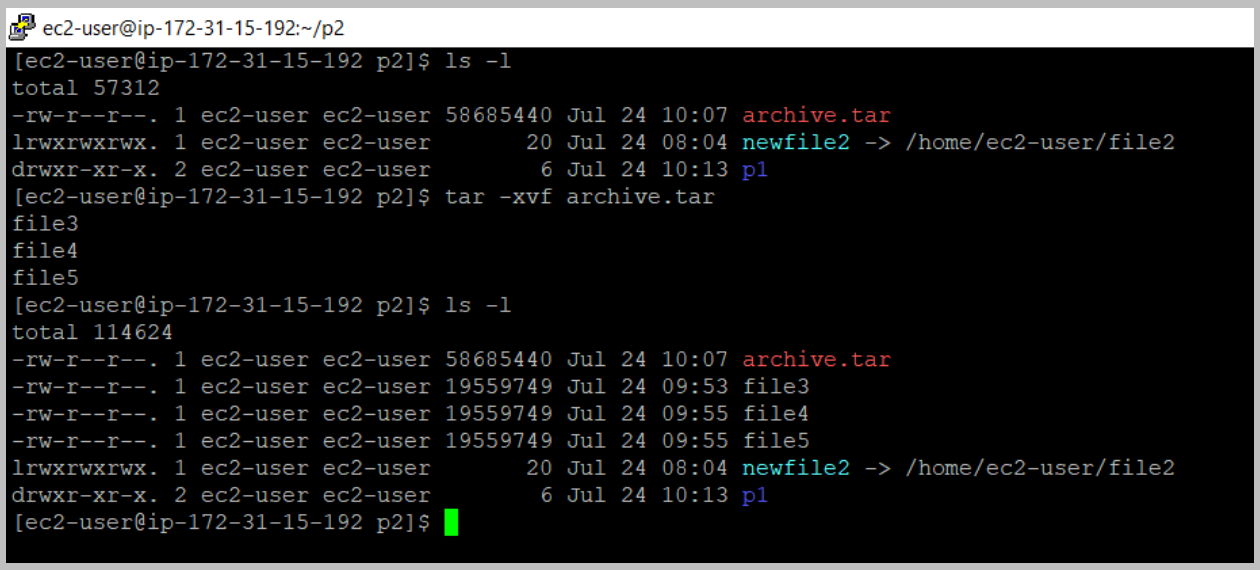
-rw-r--r-- ec2-user/ec2-user 19559749 2024-07-24 09:55 file4

-rw-r--r-- ec2-user/ec2-user 19559749 2024-07-24 09:55 file5

c. extract the content of the file using -x:

[ec2-user@ip-172-31-15-192 p2]$ tar -xvf archive.tar

This will extract the content in the current folder only.



But if you want to extract the archive folder content to some other directory then there are two ways:

a. **Approach 1:**
   Here in this approach, you need to first go to the destination folder in which you need to extract the content and then use tar command as shown in below image for example here our destination folder is “/home/ec2-user/p2/p1” and archive file present in “/home/ec2-user/p2”

```
ec2-user@ip-172-31-15-192:~/p2
[ec2-user@ip-172-31-15-192 p2]$ ls -l
total 57312
-rw-r--r--. 1 ec2-user ec2-user 58685440 Jul 24 10:07 archive.tar
lrwxrwxrwx. 1 ec2-user ec2-user       20 Jul 24 08:04 newfile2 -> /home/ec2-user/file2
drwxr-xr-x. 2 ec2-user ec2-user        6 Jul 24 10:13 p1
[ec2-user@ip-172-31-15-192 p2]$ pwd
/home/ec2-user/p2
[ec2-user@ip-172-31-15-192 p2]$ ls p1/
[ec2-user@ip-172-31-15-192 p2]$ cd p1/
[ec2-user@ip-172-31-15-192 p1]$ tar -xvf ../archive.tar
file3
file4
file5
[ec2-user@ip-172-31-15-192 p1]$ ls -l
total 57312
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:53 file3
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:55 file5
[ec2-user@ip-172-31-15-192 p1]$ pwd
/home/ec2-user/p2/p1
[ec2-user@ip-172-31-15-192 p1]$ cd ..
[ec2-user@ip-172-31-15-192 p2]$ ls -l
total 57312
-rw-r--r--. 1 ec2-user ec2-user 58685440 Jul 24 10:07 archive.tar
lrwxrwxrwx. 1 ec2-user ec2-user       20 Jul 24 08:04 newfile2 -> /home/ec2-user/file2
drwxr-xr-x. 2 ec2-user ec2-user       45 Jul 24 10:57 p1
[ec2-user@ip-172-31-15-192 p2]$ pwd
/home/ec2-user/p2
[ec2-user@ip-172-31-15-192 p2]$
```

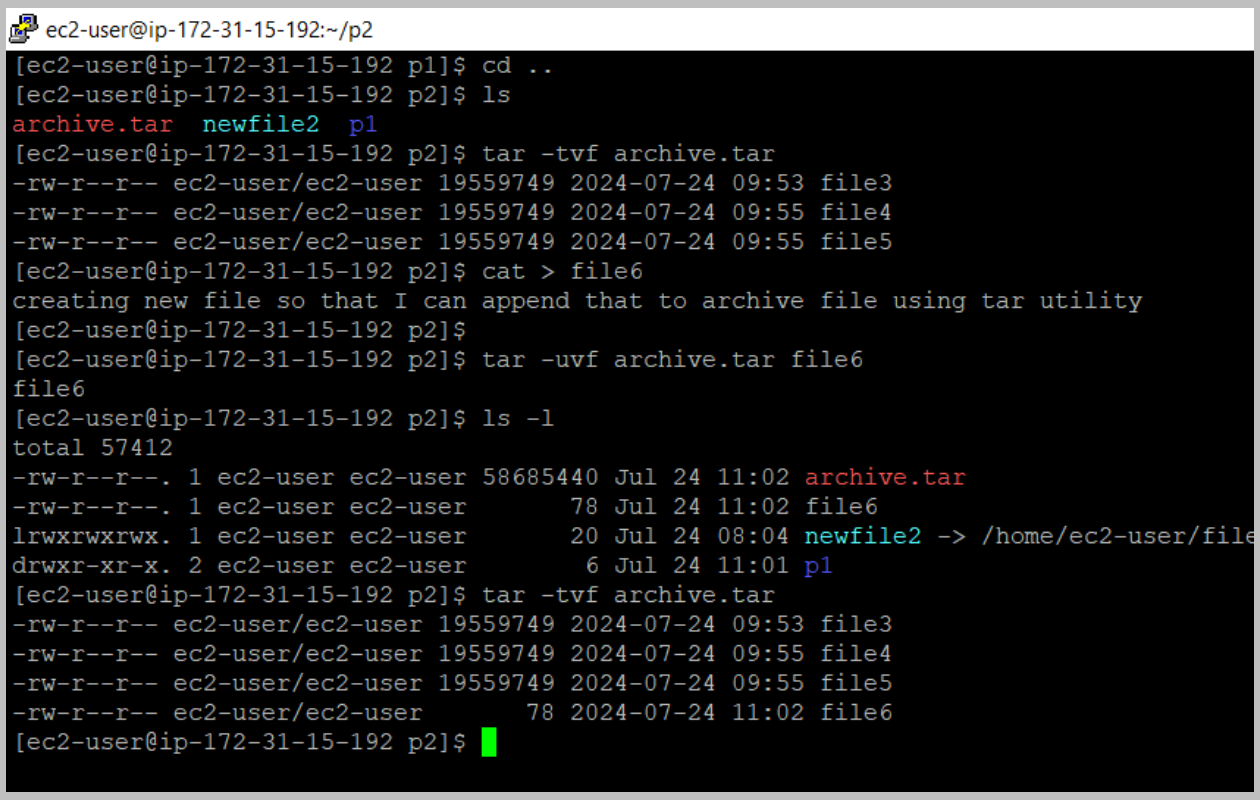
b. Approach 2:

Here we use -C option to specify the destination folder to extract the content

```
ec2-user@ip-172-31-15-192:~/p2/p1
[ec2-user@ip-172-31-15-192 p2]$ ls -l
total 57312
-rw-r--r--. 1 ec2-user ec2-user 58685440 Jul 24 10:07 archive.tar
lrwxrwxrwx. 1 ec2-user ec2-user       20 Jul 24 08:04 newfile2 -> /home/ec2-user/file2
drwxr-xr-x. 2 ec2-user ec2-user       45 Jul 24 10:57 p1
[ec2-user@ip-172-31-15-192 p2]$ rm -rf p1/*
[ec2-user@ip-172-31-15-192 p2]$ ls p1/
[ec2-user@ip-172-31-15-192 p2]$ pwd
/home/ec2-user/p2
[ec2-user@ip-172-31-15-192 p2]$ cd ..
[ec2-user@ip-172-31-15-192 ~]$ pw
-bash: pw: command not found
[ec2-user@ip-172-31-15-192 ~]$ pwd
/home/ec2-user
[ec2-user@ip-172-31-15-192 ~]$ tar -xvf /home/ec2-user/p2/archive.tar -C /home/ec2-user/p2/p1/
file3
file4
file5
[ec2-user@ip-172-31-15-192 ~]$ cd /home/ec2-user/p2/p1/
[ec2-user@ip-172-31-15-192 p1]$ ls -l
total 57312
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:53 file3
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 09:55 file5
[ec2-user@ip-172-31-15-192 p1]$
```

Note: if you do not use -f option in all above command then it will throw us an error and it won't be able to guess the archive file name.

c. Append new files to the existing archive file using -u option



Gzip Command in Linux

Gzip is one of the most popular compression algorithms that allows you to reduce the size of a file and keep the original file mode, ownership, and timestamp. Gzip also refers to the .gz file format and the gzip utility, which you can use to compress and decompress files.

gzip Command Syntax as follows:

gzip [OPTION]... [FILE]...

Gzip compresses only single files and creates a compressed file for each given file. By convention, the name of a file compressed with Gzip should end with either .gz or .z.

If you want to compress multiple files or a directory into one compressed file, first you need to create a Tar archive and then compress the .tar file with Gzip. A file that ends in .tar.gz or .tgz is a Tar archive

compressed with Gzip. Gzip is mainly used to compress text files, Tar archives, and web pages. Do not use Gzip to compress images, audio, PDF documents, and other binary files, as they are already compressed. gzip can compress only regular files. Symbolic links are ignored.

Different options available in gzip:

-k, --keep keep (don't delete) input files

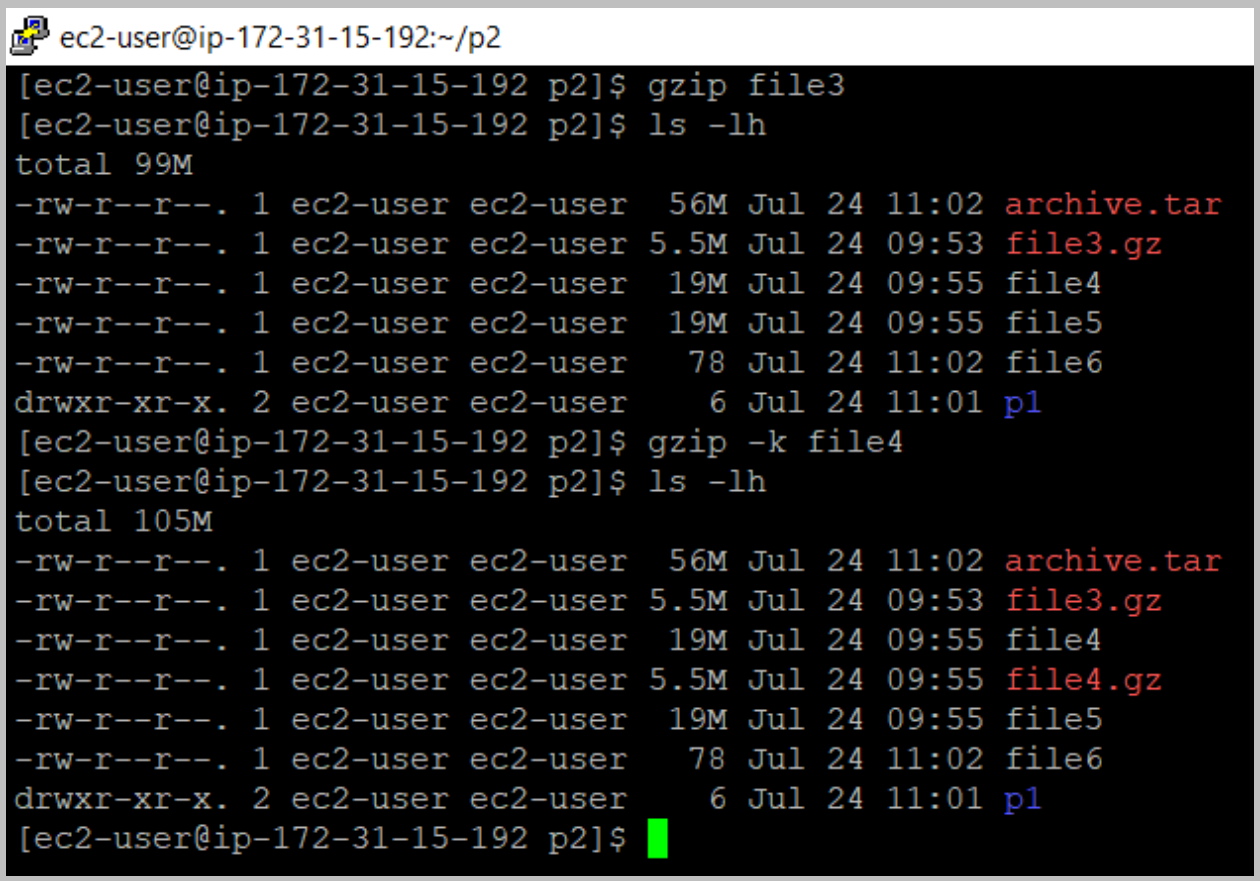
-v, --verbose verbose mode

-d, --decompress decompress

-r, --recursive operate recursively on directories

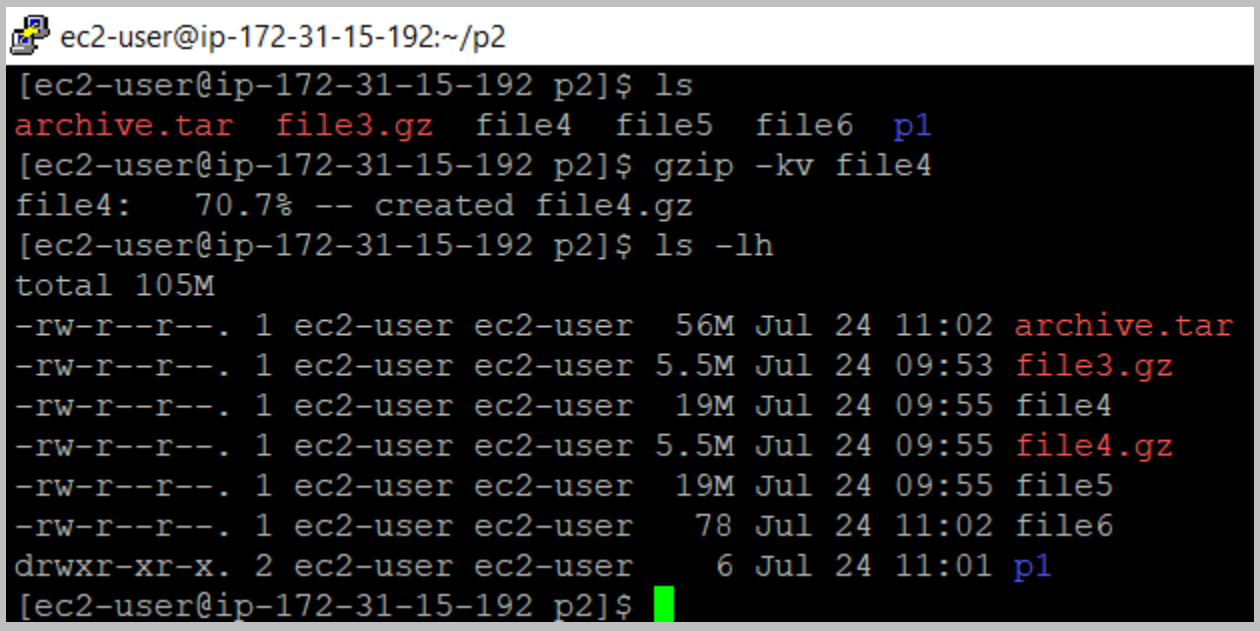
-9, --best compress better

a. As you can see in the image below we have compressed the file3 using gzip utility and it has compressed it from 19 M to 5.5M file and file3 is no longer available but if you want to keep the input file then use -k option as show in below image it has compressed file4 to file4.gzip and also kept the file4:

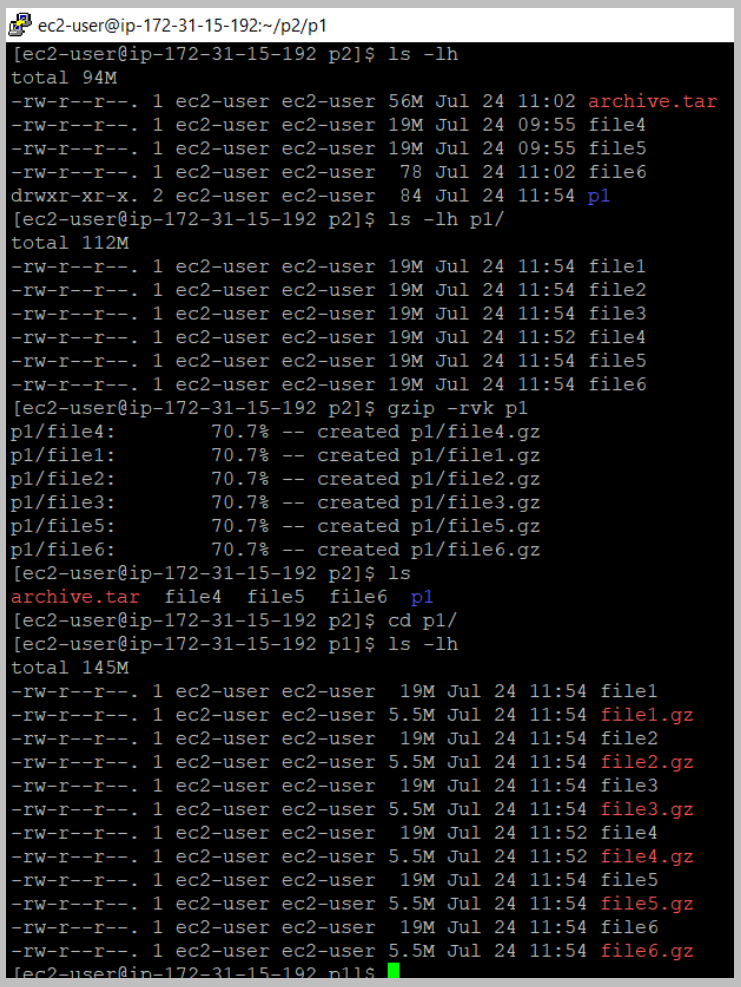


b. Verbose output

Use the -v option if you want to see the percentage reduction and the names of the files that are being processed:



c. compressing directories using gzip using -r option: as you can see in below image we tried compressing the p1 folder so it has compressed each and every file present inside p1 folder



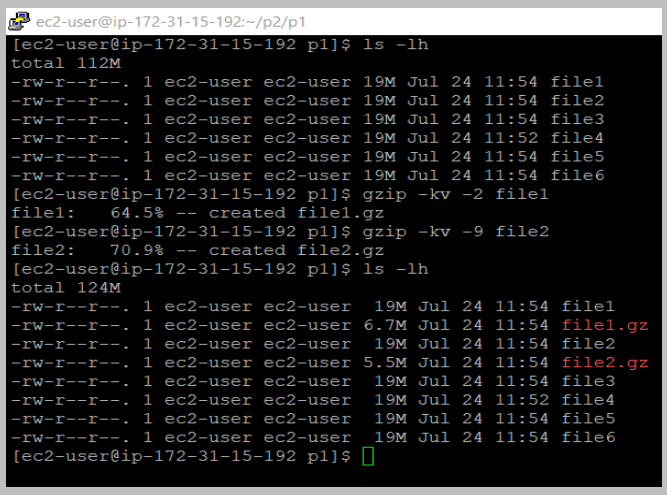
Change the compression level

gzip allows you to specify a range of compression levels, from 1 to 9. -1 or --fast means the fastest compression speed with minimal compression ratio, and -9 or --best indicates the slowest compression speed with maximum compression ratio. The default compression level is -6.

For example, to get maximum compression, you would run:

gzip -9 filename

as you can see in below image we tried compressing file1 using compression level 2 and file2 with compression level 9 so we can see the differences between the output file size of .gz:



Compressing multiple files using gzip::

A. **Approach 1:**

As you can see in below image we first tried compressing the file4 and file5 using gzip utility so it created separate compressed file of each input file. So if we need to compress multiple files using gzip then we first need to create a archive of it using tar utility and then compressed it using gzip so the resulting output file is “archiveFile4And5.tar.gz”.

```
ec2-user@ip-172-31-15-192:~/p2
[ec2-user@ip-172-31-15-192 p2]$ ls -lh
total 94M
-rw-r--r--. 1 ec2-user ec2-user 56M Jul 24 11:02 archive.tar
-rw-r--r--. 1 ec2-user ec2-user 19M Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 19M Jul 24 09:55 file5
drwxr-xr-x. 2 ec2-user ec2-user 116 Jul 24 12:03 p1
[ec2-user@ip-172-31-15-192 p2]$ gzip -kv file4 file5
file4:   70.7% -- created file4.gz
file5:   70.7% -- created file5.gz
[ec2-user@ip-172-31-15-192 p2]$ ls -lh
total 105M
-rw-r--r--. 1 ec2-user ec2-user  56M Jul 24 11:02 archive.tar
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file4.gz
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file5
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file5.gz
drwxr-xr-x. 2 ec2-user ec2-user  116 Jul 24 12:03 p1
[ec2-user@ip-172-31-15-192 p2]$ tar -cvf archiveFile4And5.tar
tar: Cowardly refusing to create an empty archive
Try 'tar --help' or 'tar --usage' for more information.
[ec2-user@ip-172-31-15-192 p2]$ tar -cvf archiveFile4And5.tar file4 file5
file4
file5
[ec2-user@ip-172-31-15-192 p2]$ ls -lh
total 142M
-rw-r--r--. 1 ec2-user ec2-user  56M Jul 24 11:02 archive.tar
-rw-r--r--. 1 ec2-user ec2-user  38M Jul 24 12:06 archiveFile4And5.tar
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file4.gz
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file5
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file5.gz
drwxr-xr-x. 2 ec2-user ec2-user  116 Jul 24 12:03 p1
[ec2-user@ip-172-31-15-192 p2]$ gzip -kv archiveFile4And5.tar
archiveFile4And5.tar:    70.7% -- created archiveFile4And5.tar.gz
[ec2-user@ip-172-31-15-192 p2]$ ls -lh
total 153M
-rw-r--r--. 1 ec2-user ec2-user  56M Jul 24 11:02 archive.tar
-rw-r--r--. 1 ec2-user ec2-user  38M Jul 24 12:06 archiveFile4And5.tar
-rw-r--r--. 1 ec2-user ec2-user  11M Jul 24 12:06 archiveFile4And5.tar.gz
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file4
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file4.gz
-rw-r--r--. 1 ec2-user ec2-user  19M Jul 24 09:55 file5
-rw-r--r--. 1 ec2-user ec2-user 5.5M Jul 24 09:55 file5.gz
drwxr-xr-x. 2 ec2-user ec2-user  116 Jul 24 12:03 p1
[ec2-user@ip-172-31-15-192 p2]$
```

B. Approach 2:

Using tar utility itself we can create gzip file using -z option in all operations

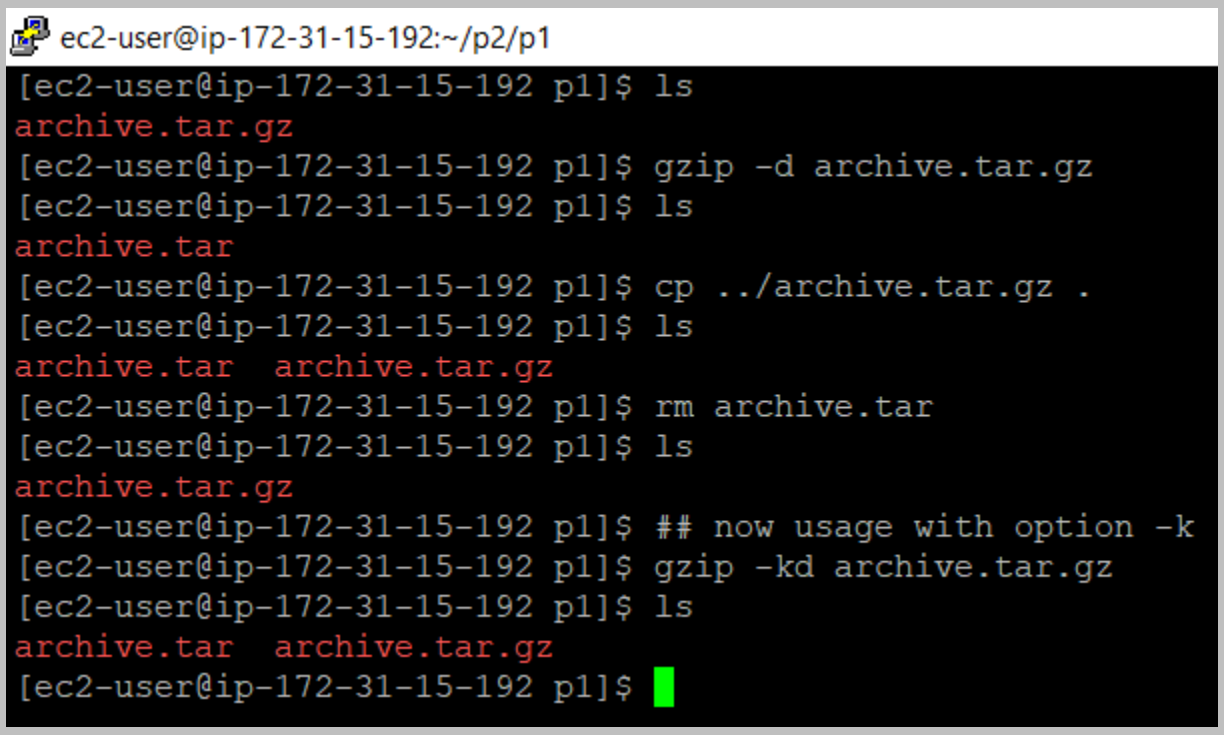
```
ec2-user@ip-172-31-15-192:~/p2
[ec2-user@ip-172-31-15-192 p2]$ ls
archive.tar  file4  file5  p1
[ec2-user@ip-172-31-15-192 p2]$ tar -zcvf archive.tar.gz archive.tar
archive.tar
[ec2-user@ip-172-31-15-192 p2]$ ls
archive.tar  archive.tar.gz  file4  file5  p1
[ec2-user@ip-172-31-15-192 p2]$
```

Extract the content of gzip file using tar utility

```
[ec2-user@ip-172-31-15-192 p2]$ tar -zxvf archive.tar.gz -C p1/
archive.tar
[ec2-user@ip-172-31-15-192 p2]$ ls p1/
archive.tar
[ec2-user@ip-172-31-15-192 p2]$
```

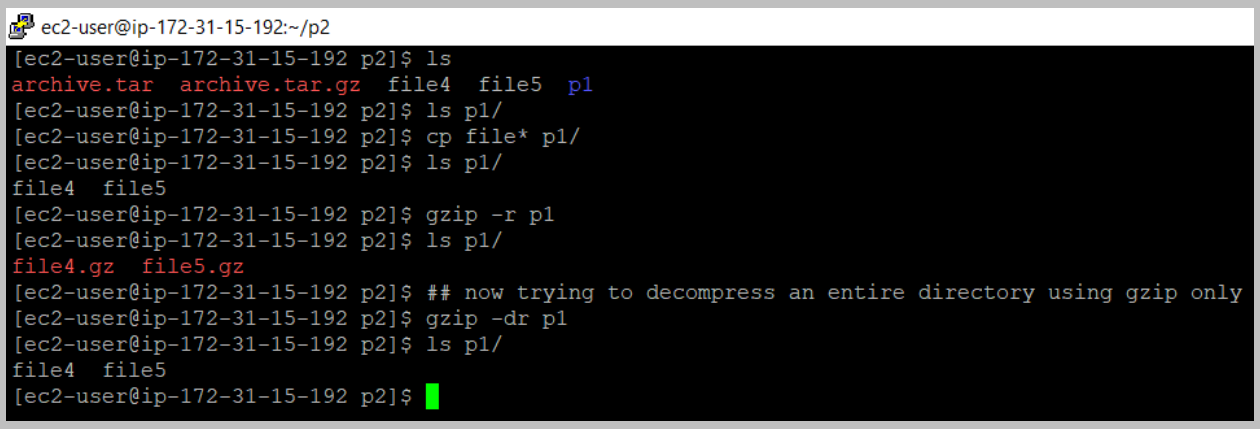
Now extract the content of .gz means gzip file using gzip utility option -d

Here we extracted the compressed .gz file using -d and with -kd (here due to -k option after extraction we can see that the archive.tar.gz file is preserved and not deleted)



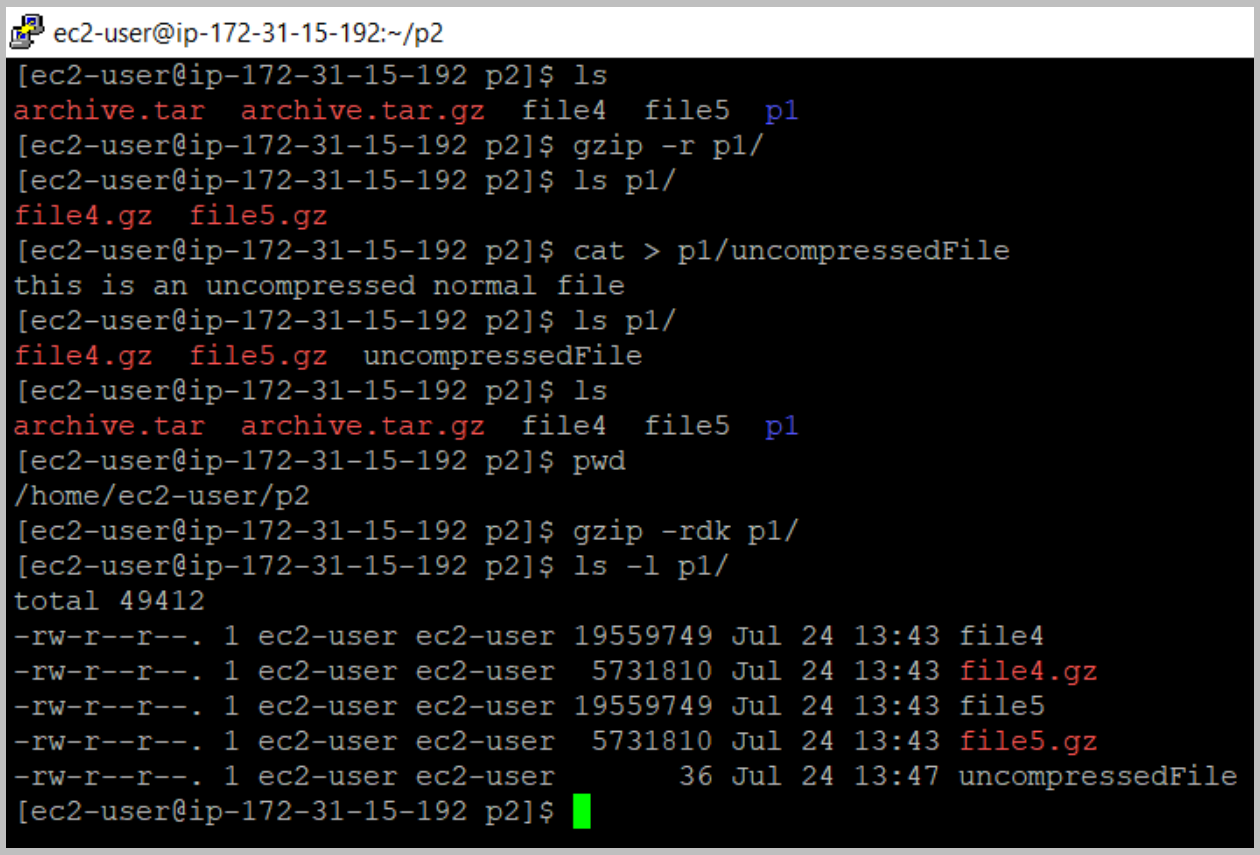
Extract entire directory using gzip utility

Here we first compressed the content of p1 directory using gzip -r p1 command and then we extracted entire directory using -dr option



Note: if you can see here in the final output after extraction the .gz file deleted that is why we didn't used -k option init

Similar to this only difference we kept some uncompressed normal files under p1 directory and when we tried extracting directory using gzip utility it has only extracted those files which are compressed as .gz



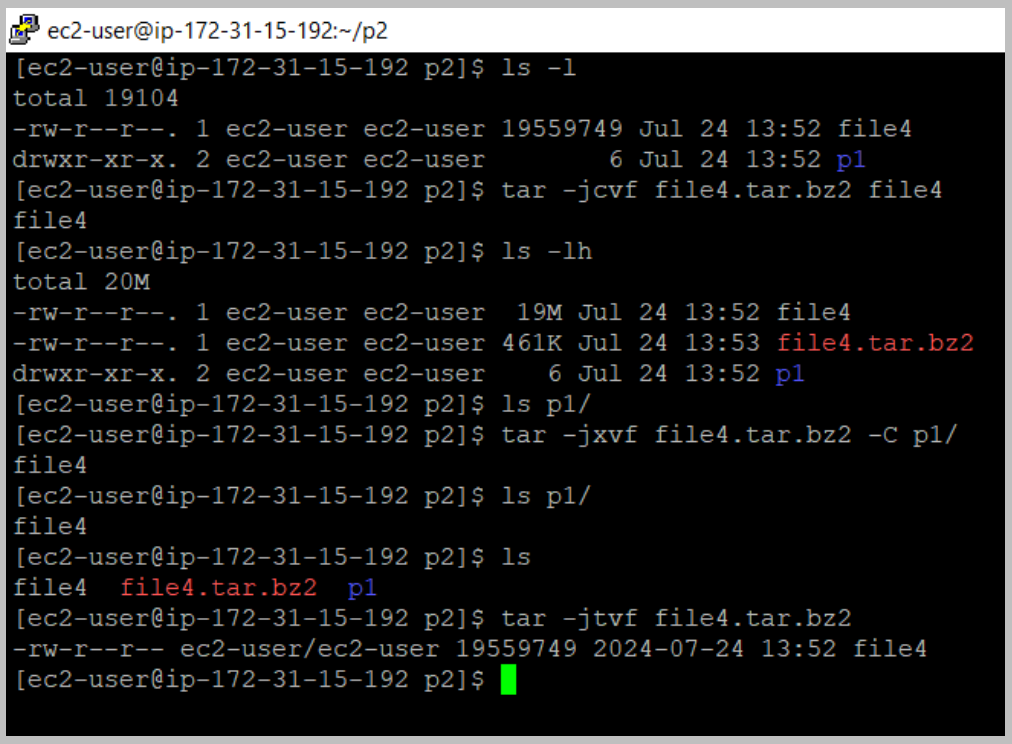
Using bzip/ bz2 utility:

The tar command allows you to create and extract tar archives. It supports a vast range of compression programs such as gzip, bzip2, lzip, lzma, lzop, xz and compress.

Bzip2 is one of the most popular algorithms for compressing tar files. By convention, the name of a tar archive compressed with bzip2 ends with either .tar.bz2 or .tbz2.

Here we can use tar utility to create, extract and view the content of bzip file

All demonstrated in below single image


ec2-user@ip-172-31-15-192:~/p2
[ec2-user@ip-172-31-15-192 p2]$ ls -l
total 19104
-rw-r--r--. 1 ec2-user ec2-user 19559749 Jul 24 13:52 file4
drwxr-xr-x. 2 ec2-user ec2-user 6 Jul 24 13:52 p1
[ec2-user@ip-172-31-15-192 p2]$ tar -jcvf file4.tar.bz2 file4
file4
[ec2-user@ip-172-31-15-192 p2]$ ls -lh
total 20M
-rw-r--r--. 1 ec2-user ec2-user 19M Jul 24 13:52 file4
-rw-r--r--. 1 ec2-user ec2-user 461K Jul 24 13:53 file4.tar.bz2
drwxr-xr-x. 2 ec2-user ec2-user 6 Jul 24 13:52 p1
[ec2-user@ip-172-31-15-192 p2]$ ls p1/
[ec2-user@ip-172-31-15-192 p2]$ tar -jxvf file4.tar.bz2 -C p1/
file4
[ec2-user@ip-172-31-15-192 p2]$ ls p1/
file4
[ec2-user@ip-172-31-15-192 p2]$ ls
file4 file4.tar.bz2 p1
[ec2-user@ip-172-31-15-192 p2]$ tar -jtvf file4.tar.bz2
-rw-r--r-- ec2-user/ec2-user 19559749 2024-07-24 13:52 file4
[ec2-user@ip-172-31-15-192 p2]$