GNU/Linux: Introduction and Administration

4h Session for the course on "Data Science. Applications to Biology and Medicine with Python and R", at IL3 - University of Barcelona[1]. April 3rd, 2024. 16:00h-19:00h.

Presentation Slides

Video recording (from a previous edition, a few years ago)

SLIDES IN PDF:
https://nextcloud.seeds4c.org/index.php/s/dszxCapKqcFR9dA[2]

Hands-on Exercise

Source data derived from data obtained from here: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee[3]

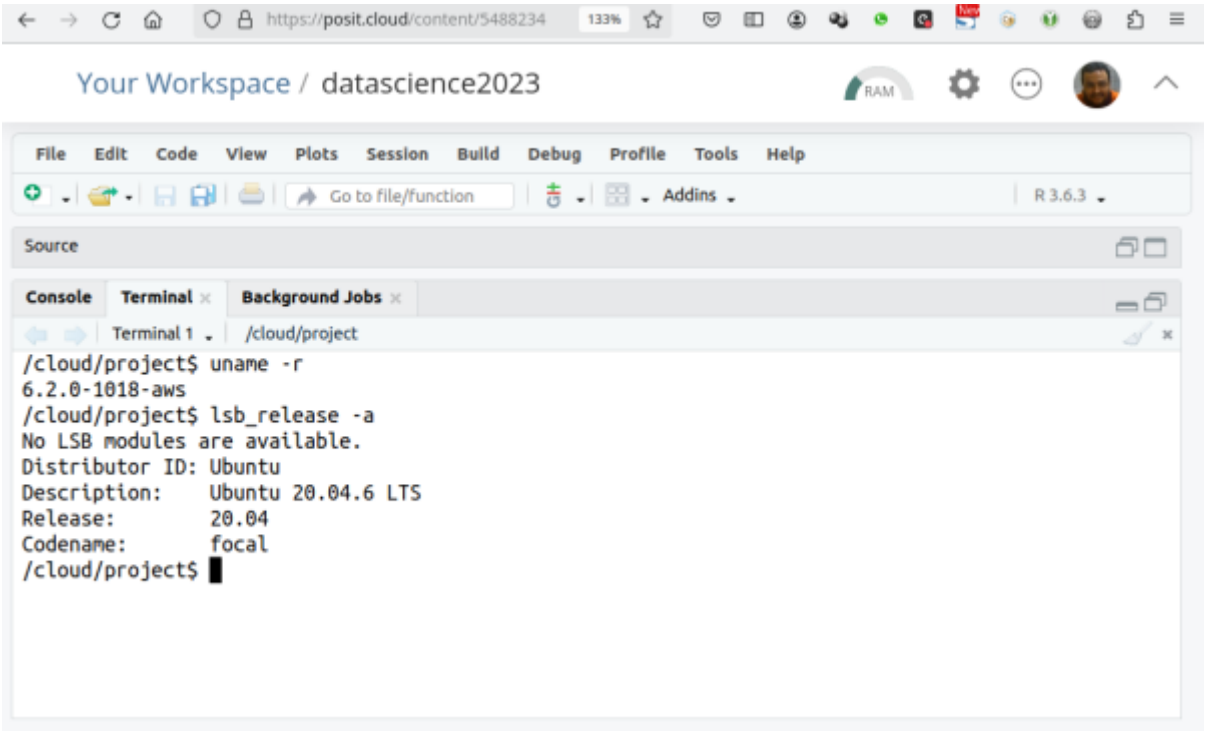
Steps:

1. **PART A**: Enter the GNU/Linux machine.
   Choose one option from the following 3 options below:
   1. Import the .ova file provided (explained within the session notes) in the VirtualBox program[4] in your own computer. **Keep in mind that it will take some time**: to download the ova file (7.6Gb), and alsoto import it to your Virtual Box (10 minutes or more),
      - OVA file:

        ```
        http://cloud.seeds4c.org/lubuntu_1804_64bit_v03.ova
        ```

      OR

      .
   2. Sign up at https://posit.cloud/plans/free[5] to get a free account. Connect to posit.cloud and use the terminal window from the RStudio server there.



OR

.

3. Connect (by means of **ssh terminal** - using Putty in Windows[6]), for instance
   [+]
   you can use usernames starting from user01, user02, user03.... user20.

   Command in a terminal

   ```
   ssh userNN@datascience.seeds4c.org
   ```

.

2. **PART B**: Fetch and subset data
   Obtain a subset of columns and rows from a dataset, using Linux simple commands in a terminal (using shell commands, not R nor Python in this case),
   1. Copy the source data file (**data_smc.csv.bz2** from the usb disk provided by the course professor), or from here for instance:
      http://cloud.seeds4c.org/data_smc.csv.bz2[8] (50Mb file, 10.000.000 rows csv file, bz2 compressed)

      Open a Linux terminal in your home folder /home/datascience/

      ```
      cd /home/userNN/ # just in case, change directory to your home folder
      wget http://cloud.seeds4c.org/data_smc.csv.bz2 # fetch the file from the internet
      ```

   2. Uncompress (`bunzip2 file.bz2 -k`) and show (with `cat file`), or use +-bzcat file.bz2 -

k+- to send to standard output (stdout) on-the-fly while keeping the source compressed file (-k)

```
bunzip2 data_smc.csv.bz2 -k
```

3. filter (keep) the first 100 rows (with `head -n100 file`)
4. save as new file: `file.csv`

Oneliner with the previous commands piped one after the other in the same line

```
bzcat data_smc.csv.bz2 -k | head -n100 > file_all.csv
```

5. filter out one column, for instance, remove column 7 (variable _), with `cut`

```
cut --complement -d',' -f7 file_all.csv > file.csv
```

6. save in zip

```
zip file.csv.zip file.csv
```

7. Change permissions so that only your user can read and write it

```
chmod 600 file.csv.zip
```

.

3. **PART C**: Expose dataset freely through webserver, for those with root access at the linux machine (option 1, with VirtualBox, from the 3 options in **PART A** above)
   1. Install Apache web server.

```
sudo apt update
sudo apt install apache2
```

      - Check that it's installed by visiting with your browser inside the virtual machine: http://localhost/
   2. Move the produced file.csv.zip to /var/www/html/ while appending the number NN fromthe username you took for the connection to the server:

```
```

```
sudo cp /home/userNN/file.csv.zip /var/www/html/fileNN.csv.zip
```

- Check if you can download it already by means of attempting to fetch the url http://localhost/fileNN.csv.zip

```
wget http://localhost/fileNN.csv.zip
```

3. change owner of that file to www-data:www-data so that it can be viewed (and downloaded) onlilne through your browser

```
sudo chown www-data:www-data /var/www/html/fileNN.csv.zip
```

Check again if you can download it (try to fetch again the url http://localhost/fileNN.csv.zip)
+

```
wget http://localhost/fileNN.csv.zip
```

That should be it: your file should be downloaded in the terminal window from the web server with the local address.
From the internet, you should be able to fecth it also at the url:

- http://datascience.seeds4c.org/fileNN.csv.zip[9]

Done!

Additional info

If you want to keep practising and learning, beyond this course session, you can do so for instance here:

1. https://davidadrian.cc/definitive-data-scientist-setup/[10]

Alias names for this page:
GNULinuxOS24 | LinuxDataScience24

[1] https://www.il3.ub.edu

[2] https://nextcloud.seeds4c.org/index.php/s/dszxCapKqcFR9dA

[3] https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Dades-meteorol-giques-de-la-XEMA/nzvn-apee

[4] https://www.virtualbox.org/wiki/Downloads

[5] https://posit.cloud/plans/free

[6] https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html

[7] https://wiki.x2go.org/doku.php/doc:installation:x2goclient

[8] http://cloud.seeds4c.org/data_smc.csv.bz2

[9] http://datascience.seeds4c.org/fileNN.csv.zip

[10] https://davidadrian.cc/definitive-data-scientist-setup/