Notes on processing large amounts of Earth Observation data

Alber Sánchez alber.ipia@inpe.br
Guilherme Mataveli

Research assistant - TreesLab
National Institute for Space Research - INPE
Brazil

August 5, 2024

Overview I

Introduction.

Computing concepts.

Hardware

- Processor.
- Memory.
- Disc (InputOutput, I/O).
- Graphic User Interface (GUI).
- Console, terminal, shell, tty.
- Client, server.

Software

Map - Reduce

Resources

- Computer Science by Crash Course.

Linux.

Basic Bash commands I

Command	Explanation
`whoami`	What is the current user name (Who am I?).
`pwd`	Print working directory (Where am I?).
`ls`	List directory contents.
`cd`	Change working directory.
`man`	Display manual pages.
`help`	Display information about commands.
`apropos`	Search the manuals and descriptions.
`rm`	Remove files or directories.
`cp`	Copy files or directories.
`mv`	Move (rename) files.
`mkdir`	Make directories.

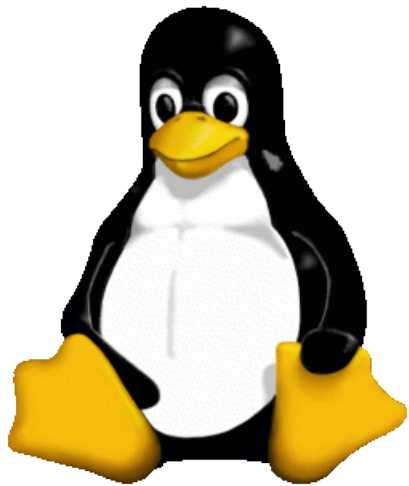
Basic Bash commands II

`less`	Open a file for reading.
`nano`	Simple text editor.
`vim`	Text editor.
`vimtutor`	The Vim tutor.
`wget`	Download data from the Internet.
`rsync`	Backup and file copying tool.
`ssh`	Remote login client.
`scp`	Secure file copy.

Basic Bash commands III

Resources

- Webpages: Linux Journey.
- Books:
  - Linux basics for Hackers [6].
  - The Linux Command Line [7].
  - Unix and Linux System Administration Handbook [3].
- Courses: Linux Foundation (LFS101).
- Videos:
  - Linux Commands for Beginners by Learn Linux TV.
  - Linux Crash Course by Learn Linux TV.
  - The Linux Command Line Ultimate Tutorial by Average Linux User.



Bash scripting.

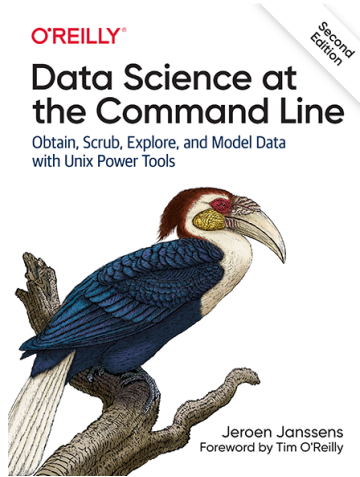
Resources

- Books: The Linux Command Line [7].
- Tutorials: The Unix Shell by Software Carpentry.
- Videos: Bash Scripting on Linux by Learn Linux TV.

How to process data without a GUI.

Data Science at the Command Line[2]

- Get data from websites, APIs, databases, and spreadsheets; scrub text, CSV, HTML, XML, JSON.
- Explore data, compute statistics, create visualizations.
- Create your own tools, re-use code.
- Parallelize and distribute data pipelines.
- Model data with dimensionality reduction, regression, and classification algorithms.
- Command line from Python, Jupyter, R, RStudio, Apache Spark.
- Read online here.



GRASS GIS

- Geographic Resources Analysis Support System.
- Vector and raster geospatial data management, geoprocessing, spatial modelling, and visualization.
- Free, Libre, and Open Source Software.
- Docker container.
- Open Source GIS: A GRASS GIS Approach [4].

Orfeo ToolBox (OTB)

- OTB is a set of state-of-the-art remote sensing tools.
- It is FOSS (Free Open Source Software).
- Wide variety of applications available: from ortho-rectification or pansharpening, all the way to classification, SAR processing, and much more!
- www.orfeo-toolbox.org

GDAL programs/utilities

- GDAL is a translator FOSS library for raster and vector geospatial data.
- It presents a single raster abstract data model and single vector abstract data model to the calling application for all supported formats.
- It also comes with a variety of useful command line utilities for data translation and processing.
- Docker container.
- https://gdal.org/programs/index.html
- Mastering GDAL tools.

R language and environment for statistical computing.

Make script take parameters

Iteration

Recomended reading:

- Iteration, chapter 26 [11].
- Functionals, chapter 9 [10].

Parallelize code

Resources

- Tutorials: R Tutorial by w3schools.
- Books:
  - An introduction to R [8],
  - Advanced R [9].

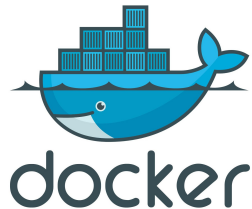
Virtualization and containerization.

Virtualization

- A host is a piece of hardware.
- A virtual machine takes pieces of host's CPU, memory, and disc and simulates new hardware.
- A software called Hypervisor takes care of virtual machines in a host.
- VirtualBox and VMware are popular virtualization software.

Containerization

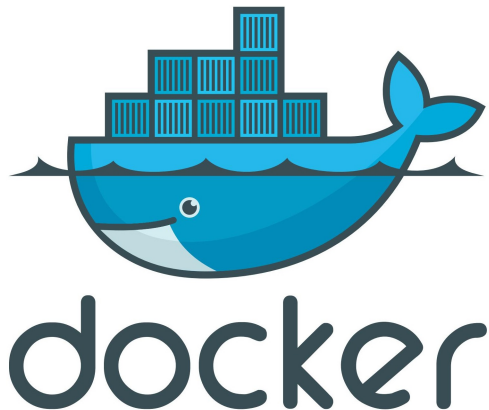
- A host is a piece of hardware.
- A virtual machine takes pieces of host's CPU, memory, and disc and simulates new hardware but it's able to share them with other host's processes.
- Reproducible, lightweight environments for processes to run.
- You're probably using them without knowing it.
- Docker and Podman are software for containerization.



podman

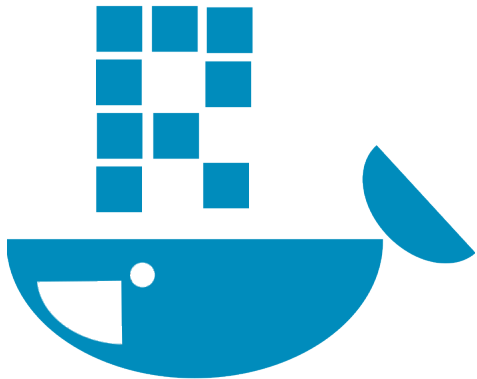
Docker

- An image is a template for creating containers.
- Images are immutable. Once created, they can't be changed.
- A container is an instance of an image.
- Docker has a repository of images called dockerhub.



Rocker project

- Docker containers for the R Environment.
- Docker images of r-base, rstudio, shiny, CUDA, geospatial.
- Portable, sandboxed, transparent, community optimized, versioned, extensible [1].
- Applications include data processing, share computing environments (HPC), teaching, packaging research reproducibility [5].
- Rocker project web page https://rocker-project.org.



Reproducible development environments

- How do I get the same environment during developing and processing?

System for Earth Observation Data Access, Processing and Analysis for Land Monitoring - SEPAL.

SEPAL

- Cloud computing-based platform for autonomous land monitoring using remote sensing data.
- This platform allows users to access powerful cloud-computing resources to query, access and process satelllite data quickly and efficiently for conducting advanced analysis.
- It is part of OpenForis.

Setup

Create accounts at:

- SEPAL.
- Google Earth Engine.
- Collect Earth Engine.
- NICFI-PlanetLab data.

Get data

- Connect SEPAL to Google Earth Engine.
- Connect SEPAL to NICFI-PlanetLab (Norway's International Climate and Forest Initiative).

Recipes

- Quickly and efficiently query and process satellite data.
- A recipe is a record of steps and parameters used to make a data set (e.g. classification).
- Access GEE imagery catalog and run planetary-scale analysis without a line of code.
- Recipes include creating mosaics (optical, radar, planet), supervised classifications (of images or time series).

Modules

- GIS tools that complement SEPAL's recipes.
- Based on advanced GIS libraries.

Workflows

- Combinations of Recipes, modules, and tools to perform complex data analysis.

CLI utilities

- Command Line Interface utilities.
- GDAL, Google Drive, GEE, GuidosToolbox Workbench, Open Foris Geospatial Toolbox, Orfeo Toolbox, Python, R code.
- IDEs.

IDEs

- Integrated Development Environments.
- JupyterLab.
- Jupyter Notebook.
- RStudio.

Start a machine

Get data

Start an R-spatial container

Get root password

Test case: sepal.io

Script for...

- Suppose we have an R script for computing...

Take home message

References I

[1] Carl Boettiger and Dirk Eddelbuettel. "An Introduction to Rocker: Docker Containers for R". In: *The R Journal* 9.2 (2017), p. 527. ISSN: 2073-4859. DOI: 10.32614/RJ-2017-065. (Visited on 07/31/2024).

[2] Jeroen Janssens. *Data Science at the Command Line: Obtain, Scrub, Explore, and Model Data with Unix Power Tools*. Second edition. Sebastopol, CA: O'Reilly Media, Inc., 2021. ISBN: 978-1-4920-8788-5.

[3] Evi Nemeth Garth Snyder Trent Hein R. Ben Whaley Dan Mackin. *UNIX and Linux System Administration Handbook, 5th Edition*. Place of publication not identified: Addison-Wesley Professional, 2017. ISBN: 978-0-13-427830-8.

[4] Markus Neteler and Helena Mitasova, eds. *Open Source GIS*. Boston, MA: Springer US, 2008. ISBN: 978-0-387-35767-6. DOI: 10.1007/978-0-387-68574-8. (Visited on 03/08/2024).

References II

[5] Daniel Nüst et al. "The Rockerverse: Packages and Applications for Containerisation with R". In: *The R Journal* 12.1 (2020), p. 437. ISSN: 2073-4859. DOI: 10.32614/RJ-2020-007. (Visited on 07/31/2024).

[6] OccupyTheWeb. *Linux Basics for Hackers: Getting Started with Networking, Scripting, and Security in Kali*. First edition. San Francisco: No Starch Press, Inc, 2018. ISBN: 978-1-59327-855-7.

[7] William Shotts. *The Linux Command Line*. 5th ed. LinuxCommand.org, Jan. 2019. (Visited on 07/02/2024).

[8] William N Venables, David M Smith, et al. *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics, Version 1.9. 1*. 4.4.1. r-project.org, June 2024. (Visited on 07/03/2024).

[9] Hadley Wickham. *Advanced R*. The R Series. Boca Raton, FL: CRC Press, 2015. ISBN: 978-1-4665-8696-3. (Visited on 07/03/2024).

References III

[10] Hadley Wickham. *Advanced R*. Second edition. Chapman & Hall/CRC: The R Series. Boca Raton London New York: CRC Press, Taylor & Francis Group, 2019. ISBN: 978-0-8153-8457-1 978-0-367-25537-4.

[11] Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Second edition. Beijing ; Sebastopol, CA: O'Reilly, 2023. ISBN: 978-1-4920-9740-2.