Linux y HPC para big data

Alex Di Genova

01/04/2024

Visión global

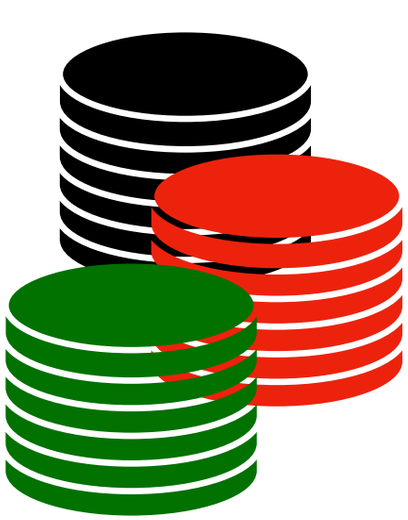
Byte	B	10^0	1
Kilobyte	KB	10^3	1,000
Megabyte	MB	10^6	1,000,000
Gigabyte	GB	10^9	1,000,000,000
Terabyte	TB	10^{12}	1,000,000,000,000
Petabyte	PB	10^{15}	1,000,000,000,000,000
Exabyte	EB	10^{18}	1,000,000,000,000,000,000

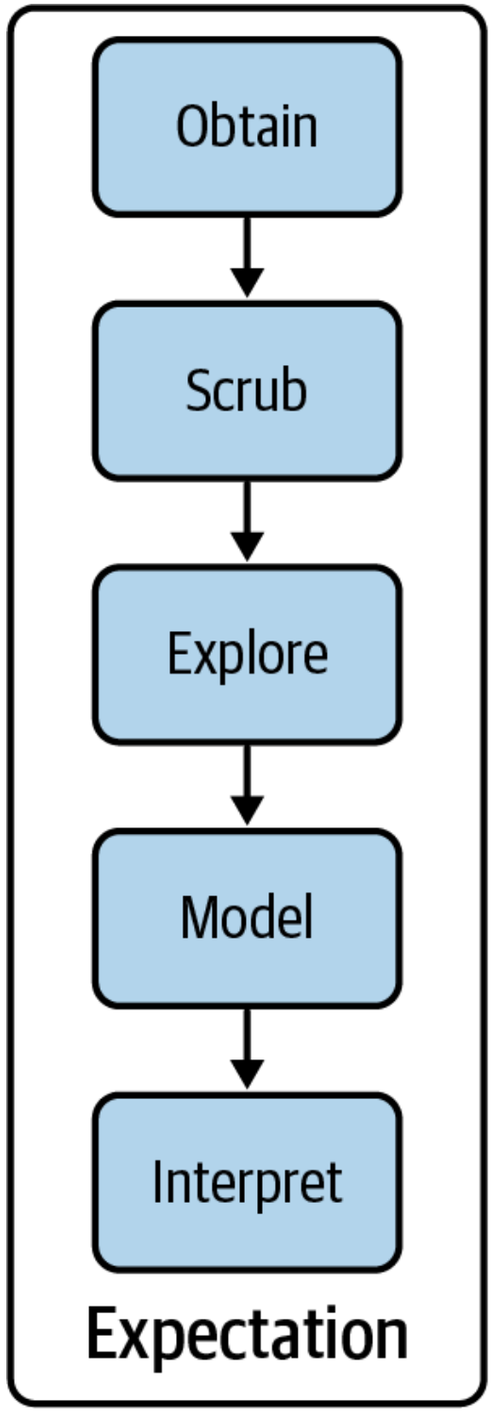
- Astronomía
- Genómica
- Redes Sociales
- Youtube

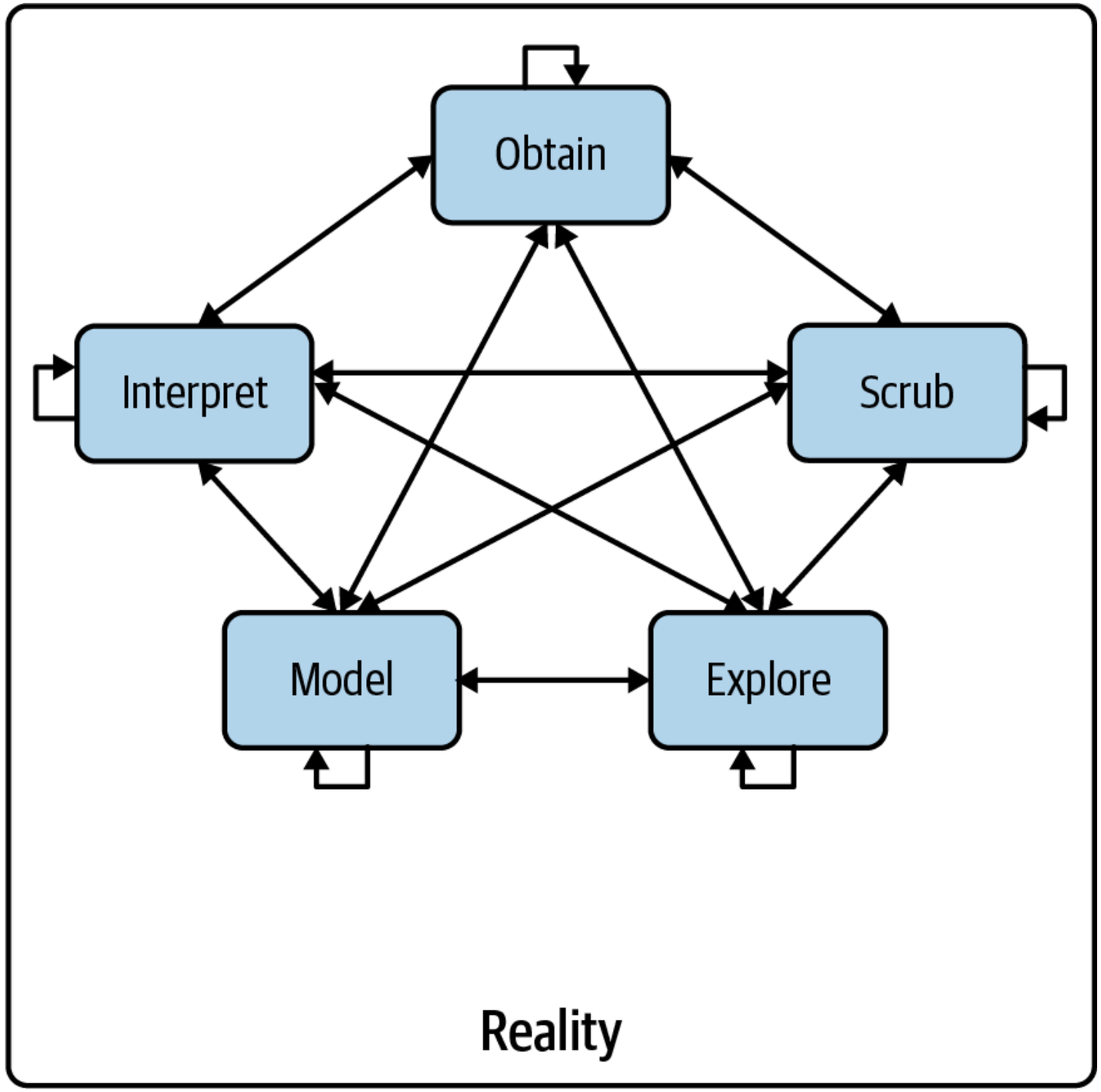
Stephens, Zachary D., et al. "**Big data: astronomical or genomical?**." *PLoS biology* 13.7 (2015): e1002195.

Data science

Expectation vs Reality





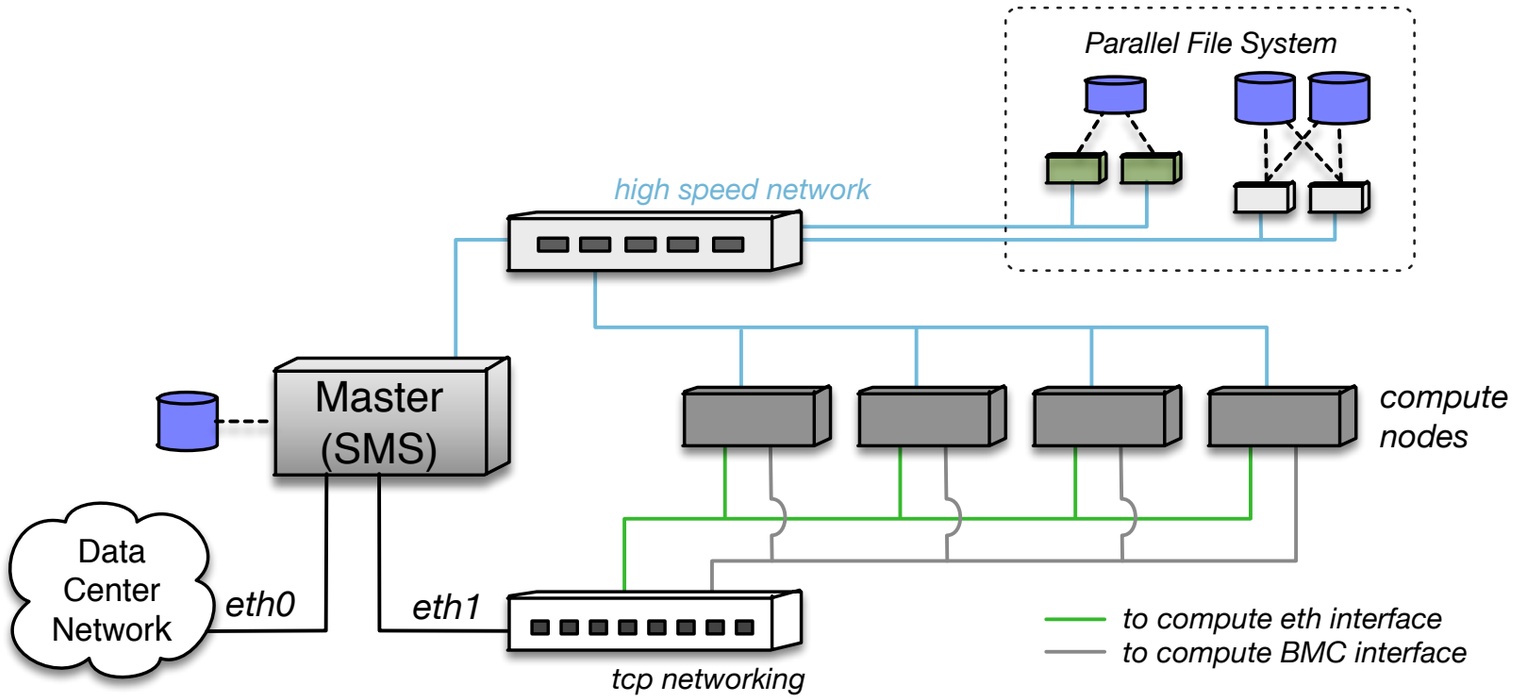


DATA -> INFORMATION

Sistemas distribuidos

Tipos

- Sistemas de cómputo distribuido
  - Clúster de cómputo
    - Hardware similar
    - Red local de alta velocidad
    - Mismo sistema operativo



The **master** handles the allocation of nodes to a particular parallel program, maintains a batch queue of submitted jobs, and provides an interface for the users of the system.

Linux

A operating system

User Processes

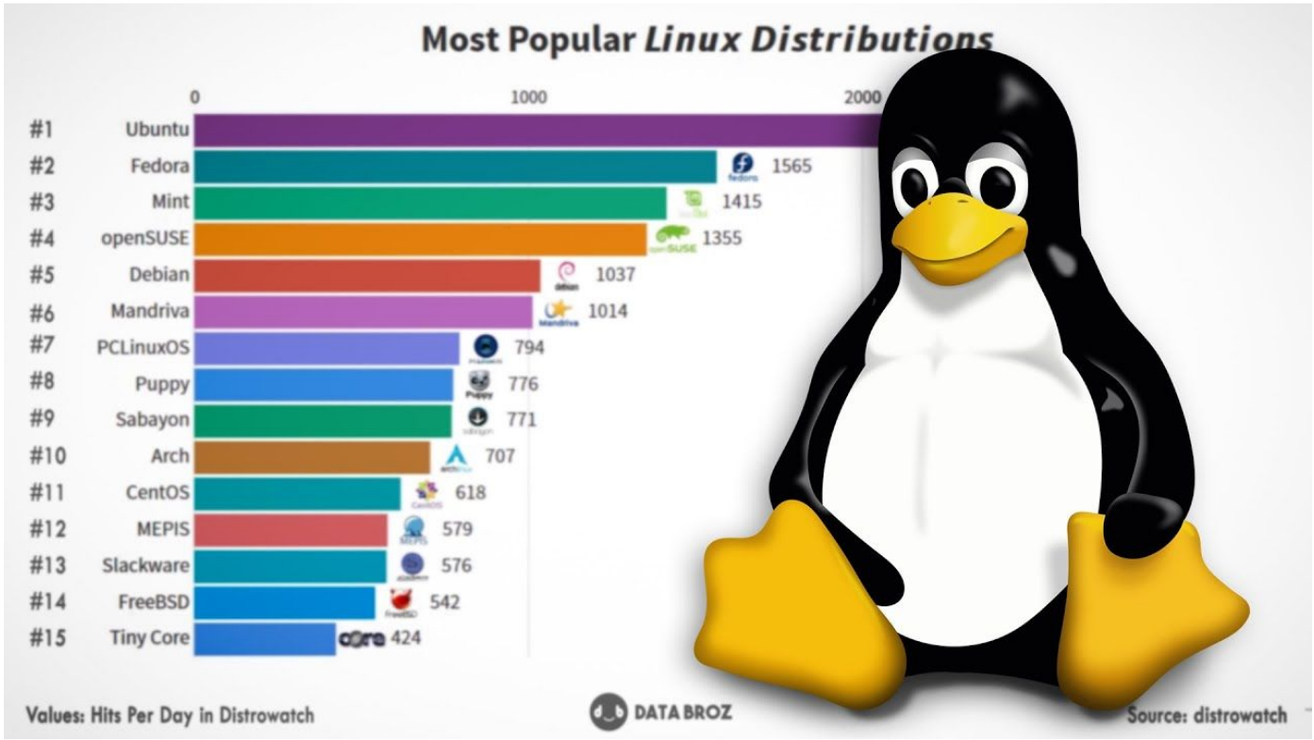
- Graphical User Interface
- Servers
- Shell

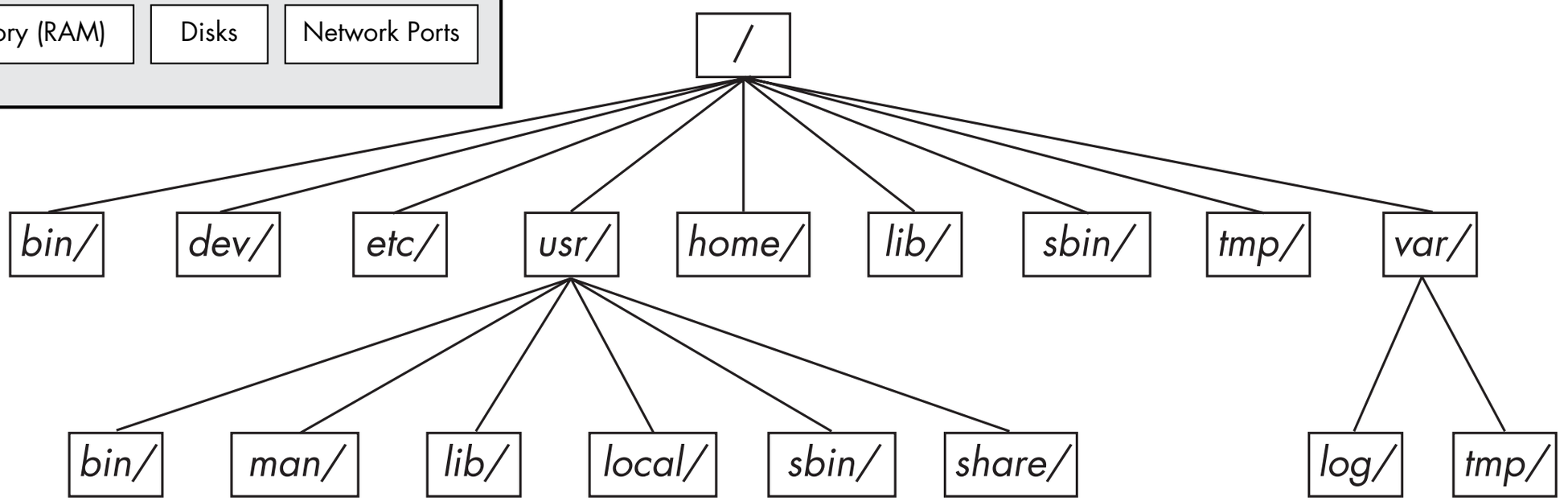
Linux Kernel

- System Calls
- Process Management
- Memory Management
- Device Drivers

Hardware

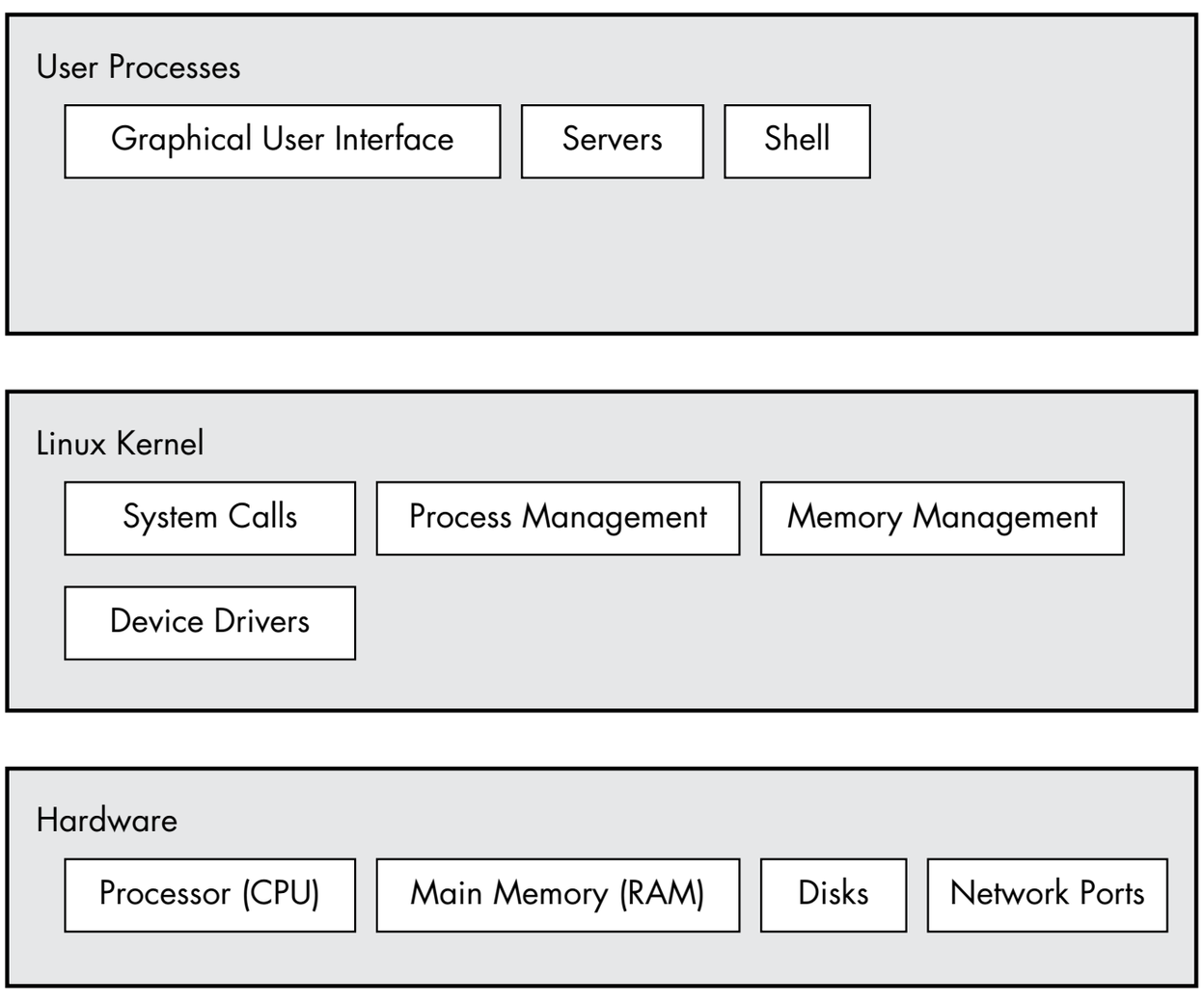
- Processor (CPU)
- Main Memory (RAM)
- Disks
- Network Ports



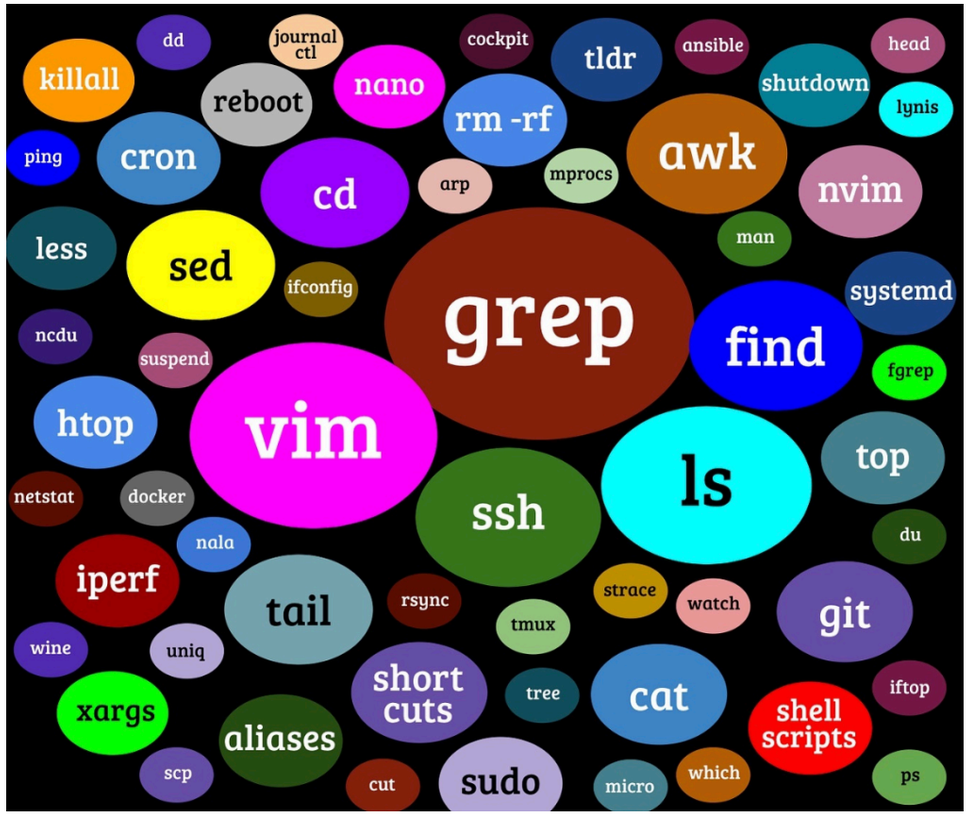


Linux

A operating system



- The *shell* is a program that runs commands.
- The *shell* also serves as a small programming environment.



- https://github.com/trinib/Linux-Bash-Commands

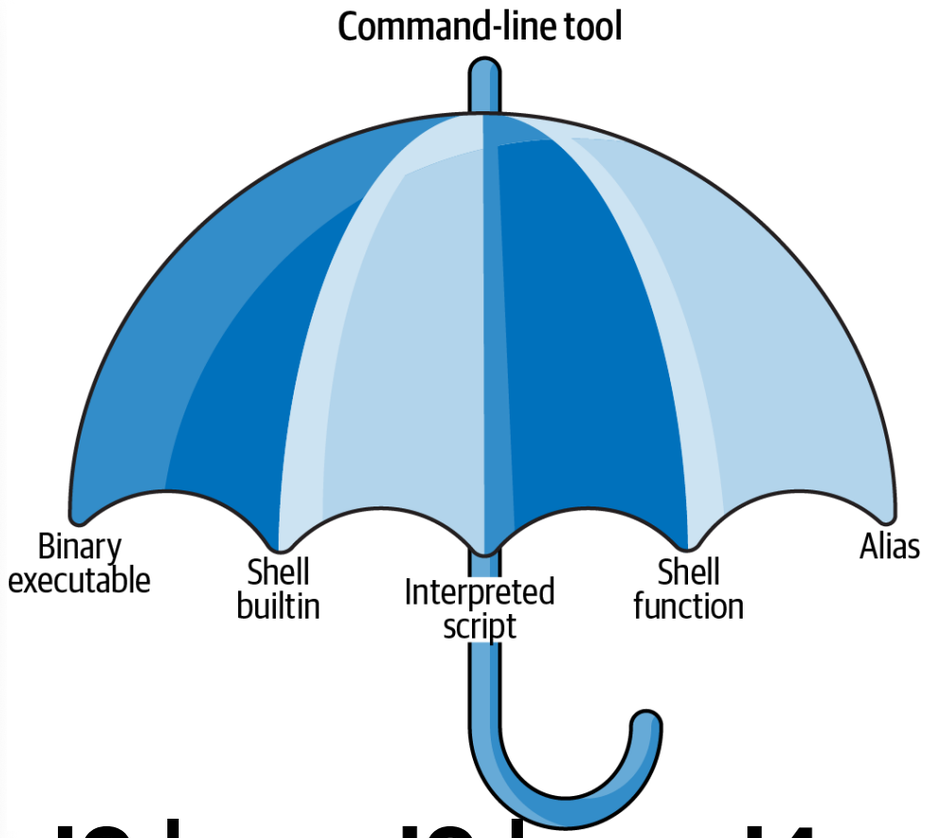
Linux

Shell command line

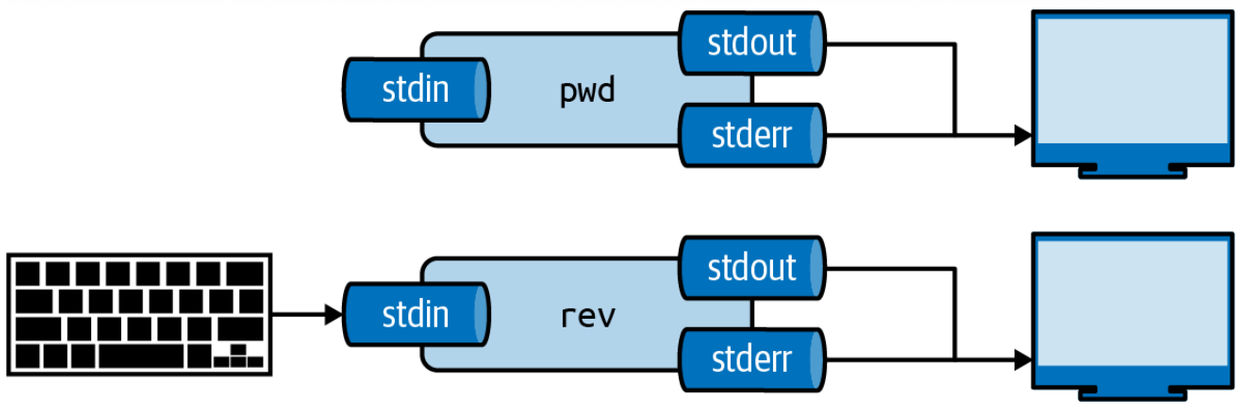
clases — adigenova@host04:~ — ssh adigenova@172.16.105.104 — 80×23

...@host04:~ — -bash ... ...R/pipelines — -bash ... ...ethylation — -bash ... ...trelka_pool — -bash ... ...ova@172.16.105.104 ...elka_pool] — -bash ... ...egfs/home — -bash ... +

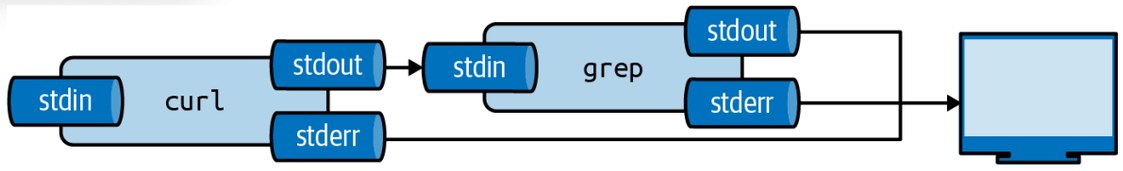
```
[adigenova@host04 ~]$ echo "The command line is the force DCBI" | cowsay -f tux
 ____________________________________
< The command line is the force DCBI >
 ------------------------------------
   \
    \
        .--.
       |o_o |
       |:_/ |
      //   \ \
     (|     | )
    /'\_   _/`\
    \___)=(___/

[adigenova@host04 ~]$
```



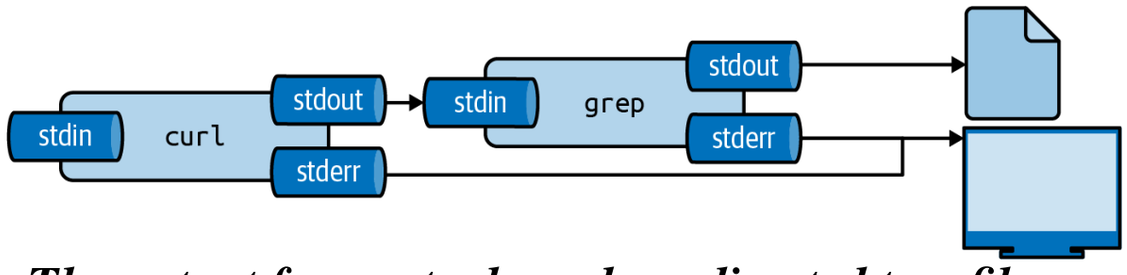
cmd1 | cmd2 | cmd3 | cmd4 ...



Every tool has three standard streams: standard input (stdin), standard output (stdout), and standard error (stderr)



The output from a tool can be piped to another tool



The output from a tool can be redirected to a file

Linux

Basic shell comands

User Management

```
sudo su                                  # Switch to root user
sudo foo                                 # Execute commands(has permission denied) as the root user
sudo nano /foo/foo.txt                   # Open directories and files(is not writable) as the root user
su username                              # Switch to a different user

passwd                                   # To change the password of a user
adduser username                         # To add a new user
userdel username                         # To remove user
userdel -rl--remove username             # To remove user with home directory and mail spool
usermod -al--append -Gl--groups GROUPNAME USERNAME # To add a user to a group
deluser USER GROUPNAME                   # To remove a user from a group

last                                     # Display information of all the users logged in
last username                            # Display information of a particular user
w                                        # Display who is online
```

System Information

```
uname -s                 # Print kernel name
uname -r                 # Print kernel release
uname -m                 # Print Architecture
uname -o                 # Print Operating System
uname -a                 # Print all Systen info

lsb_release -a           # Print distribution-specific information
dpkg --print-architecture # Print-architecture by name

cat /proc/cpuinfo        # Show cpu info
cat /proc/meminfo        # Show memory info
```

Terminal Multiplexers

Start multiple terminal sessions. Active sessions persist r

```
tmux             # Start a new session (CTRL-b + d to detach)
tmux ls          # List all sessions
tmux attach -t 0 # Reattach to a session

screen           # Start a new session (CTRL-a + d to detach)
screen -S foo    # Start a new named session
screen -ls       # List all sessions
screen -R 31166  # Reattach to a session
```

Secure Shell Protocol (SSH)

```
ssh hostname                 # Connect to hostname using your current user name (p 22)
ssh -i foo.pem hostname      # Connect to hostname using the identity file
ssh user@hostname            # Connect to hostname using the user over the default SSH port 22
ssh user@hostname -p 8765    # Connect to hostname using the user over a custom port
ssh ssh://user@hostname:8765 # Connect to hostname using the user over a custom port
```

Secure Copy

```
scp foo.txt ubuntu@hostname:/home/ubuntu          # Copy foo.txt into the specified remote directory
scp ubuntu@hostname:/home/ubuntu/foo.txt .    # Copy foo.txt from the specified remote directory
```

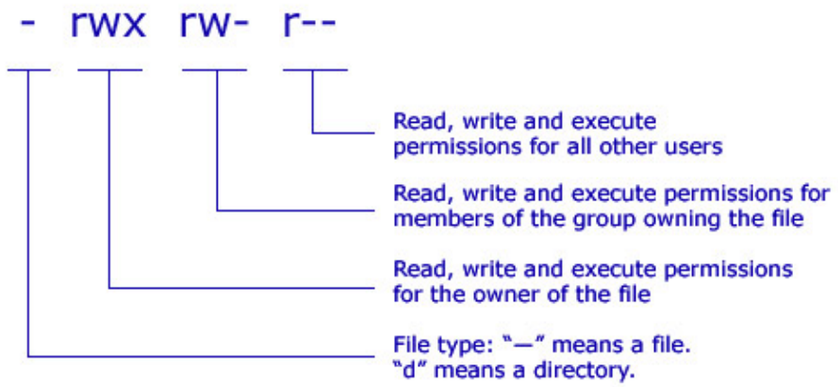
https://linuxupskillchallenge.org/

Manipulating files

Linux

Everything is a file

- File types
  - Regular files (text, images)
  - Directories (list of files)
  - Symbolic links
  - Devices and peripherals
  - Pipes
  - Sockets (inter-process communication)
- File names
  - Case sensitive, any character (/), extension not needed (but user?)
- File paths
  - Relative path: ../work/file1.txt
  - Absolute path: /home/adigenova/work/file1.txt
- File access rights
  - Types
    - Read(r), write(w), execute(x)
  - Levels
    - User (u), Group (g), Other(o)



Linux

Everything is a file

- Common text files
  - CSV -> comma-separated values
  - TSV -> Tab-separated values
  - TXT -> text file (space separated)
  - xls|xlsx -> Excel files (text or binary?)
  - SQL database

- Getting data
  - curl -LO https://burntsushi.net/stuff/worldcitiespop.csv
  - wget https://burntsushi.net/stuff/worldcitiespop.csv
  - Tar zxvf file.tar.gz, gzip , gz, zip … (compressed data)
  - Scp, cp, rsync …
- Text file editors
  - VIM/Vi
  - Nano

https://github.com/awesomedata/awesome-public-datasets

Linux

Working with text files line by line

- Viewing and Creating Files: ***cat, tac, more, less, echo, touch***
- Editing Files: ***nano, vi, vim***
- File Content Manipulation:
    - ***grep:*** Searches for patterns within files; can be used to find specific text.
    - ***sed:*** Stream editor for filtering and transforming text in a file or input from a pipeline.
    - ***awk:*** A programming language designed for text processing and manipulation.
    - ***paste:*** Merges lines of files horizontally.
    - ***cut, sort, uniq, tr, wc, head, tail.***

Linux

Working with text files line by line

- Comparing Files:
  - **diff :** Compares files line by line.
- Regular Expressions and Pattern Matching: ***grep, egrep, fgrep, sed, awk***
- Text Processing:
  - ***join:*** Joins lines of two files on a common field.
  - ***split:*** Splits a file into fixed-size pieces.
  - ***nl*** (add line numbers), ***expand*** (tab to spaces)***, unexpand*** (spaces to tab)***, fold y fmt*** (text format).
- File Encoding and Character Conversion: ***dos2unix, unix2dos***

Linux

Working with text files line by line

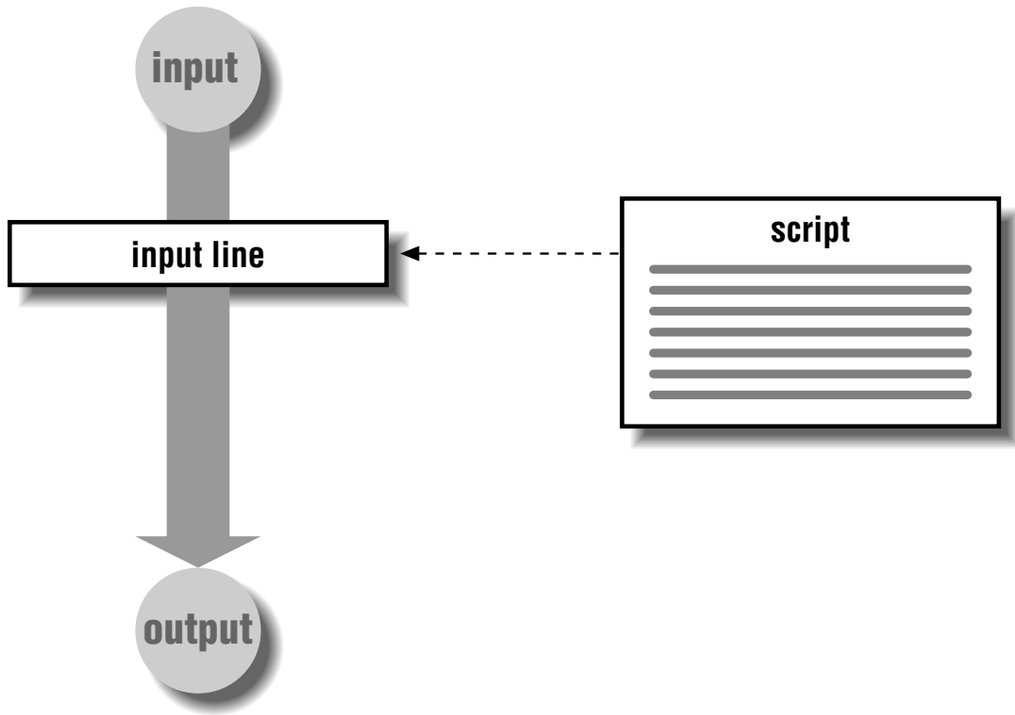
```
Country,City,AccentCity,Region,Population,Latitude,Longitude
ad,aixas,Aixàs,06,,42.4833333,1.4666667
ad,aixirivali,Aixirivali,06,,42.4666667,1.5
ad,aixirivall,Aixirivall,06,,42.4666667,1.5
ad,aixirvall,Aixirvall,06,,42.4666667,1.5
ad,aixovall,Aixovall,06,,42.4666667,1.4833333
ad,andorra,Andorra,07,,42.5,1.5166667
ad,andorra la vella,Andorra la Vella,07,20430,42.5,1.5166667
ad,andorra-vieille,Andorra-Vieille,07,,42.5,1.5166667
ad,andorre,Andorre,07,,42.5,1.5166667
```

- head worldcitiespop.csv
- wc -l worldcitiespop.csv
- head -100 worldcitiespop.csv > sample_worldcitiespop.csv
- cut -d',' -f1,5 worldcitiespop.csv > country_population.csv
- tail -n +2 worldcitiespop.csv | sort -t',' -k5 -nr | head -10
- cut -d',' -f1 worldcitiespop.csv | sort | uniq | tail -n +2
- awk -F',' ‘NR>1 && $5 > 1000000' worldcitiespop.csv
- awk -F',' 'NF > 1 {pop[$1]+=$5} END{ for(c in pop){print c" "pop[c]}}' worldcitiespop.csv | sort -rn -k2,2 | head
- Add country name using join..

Linux

Sed and Awk —- Power tools for editing files

- The basic function of awk is to search files for lines (or other units of text) that contain certain patterns.
  - A Pattern-Matching Programming Language
  - awk 'program' input-file1 input-file2 ...
  - View a text file as a textual database made up of records and fields.
  - Use variables to manipulate the database.
  - Use arithmetic and string operators.
  - Use common programming constructs such as loops and conditionals.
  - Generate formatted reports.



- Sed
  - Sed is a “non-interactive” stream-oriented editor.
  - To automate editing actions to be performed on one or more files.
  - To simplify the task of performing the same edits on multiple files.
  - To write conversion programs.

```
sed ’s/2022/2024/g’ file.txt
sed ’s/2022/2024/g ; s/casa/home/g’ file.txt

awk ‘{print $1}’ file.txt
awk ‘{print $1}’ *.txt
awk ‘/perro/ {print $0}’ *.txt
```