INTRODUCCIÓN AL PROCESAMIENTO DISTRIBUIDO DE BIG DATA

MAESTRÍA EN CIENCIA DE DATOS

Unidad 1

2. Del PC al DataLake: Operación Básica de GNU/Linux y Configuración de Red.

Gabriel Tamura
gtamura@icesi.edu.co

UNIVERSIDAD ICESI

The PC as the Data Scientist's Most Basic Tool

What
mani
their

Wher

tudy

Three ways of using your PC as the data manipulation tool:

1. Individual tool
2. Individual-virtualized tool
3. Part of a cluster tool

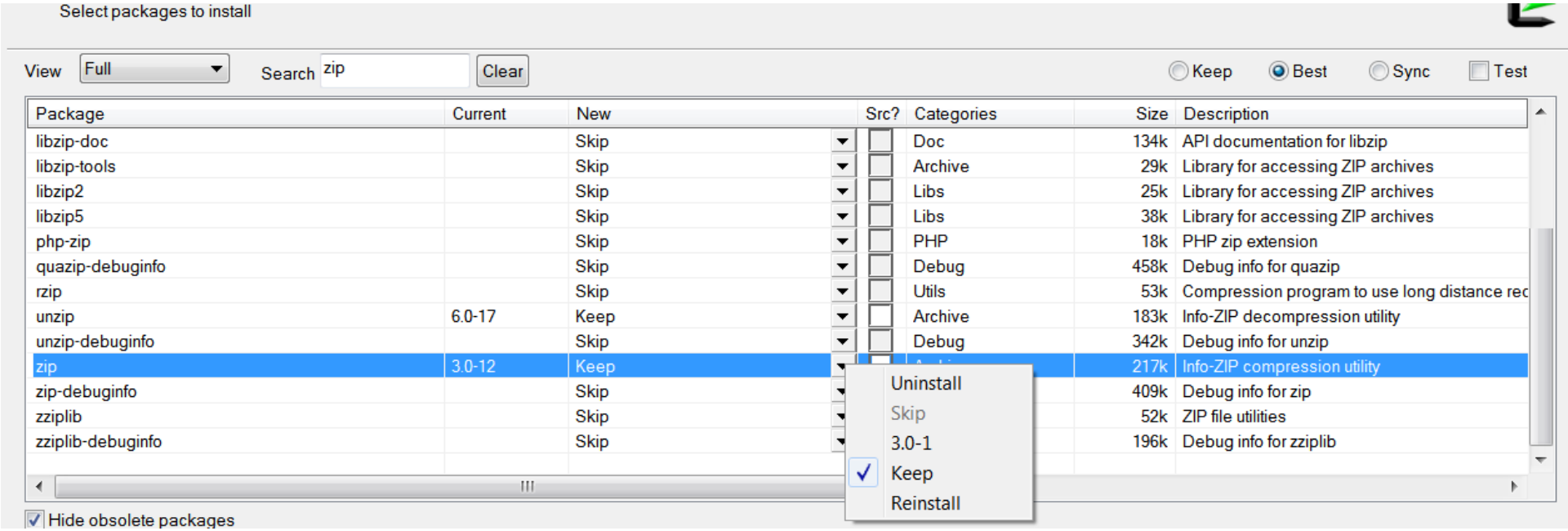
Verify the Installation of Basic Tools: CygWin for Windows

Follow the instructions for installing Cygwin:

- General description in Moodle, Unit 1
- Detailed instructions in PDF
- Check:
  - ssh / scp --version
  - wget --version
  - split --version
  - cat --version
  - sed --version
  - awk --version
  - zip --version
  - unzip --version

Verify the Installation of Basic Tools: CygWin for Windows

If zip/unzip are not found, run setup.exe again and choose a different mirror. Search for them and select the action:



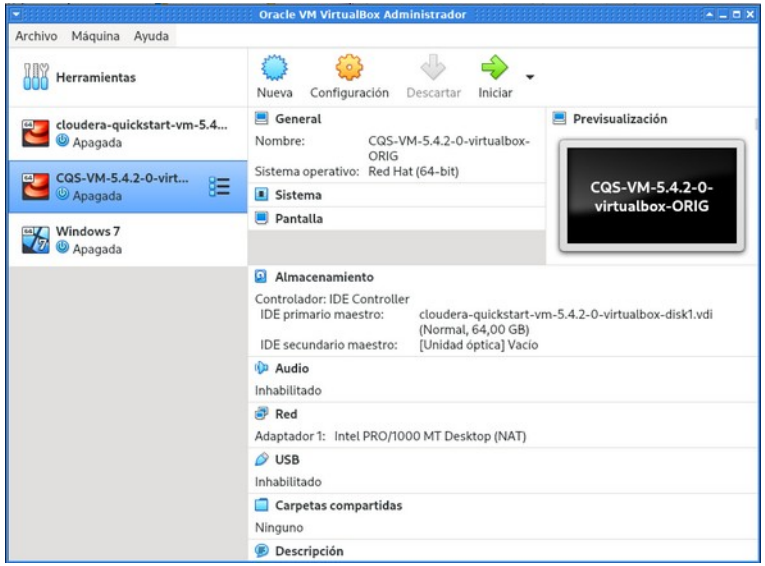
Verify the Installation of Preconfigured Virtual Machines

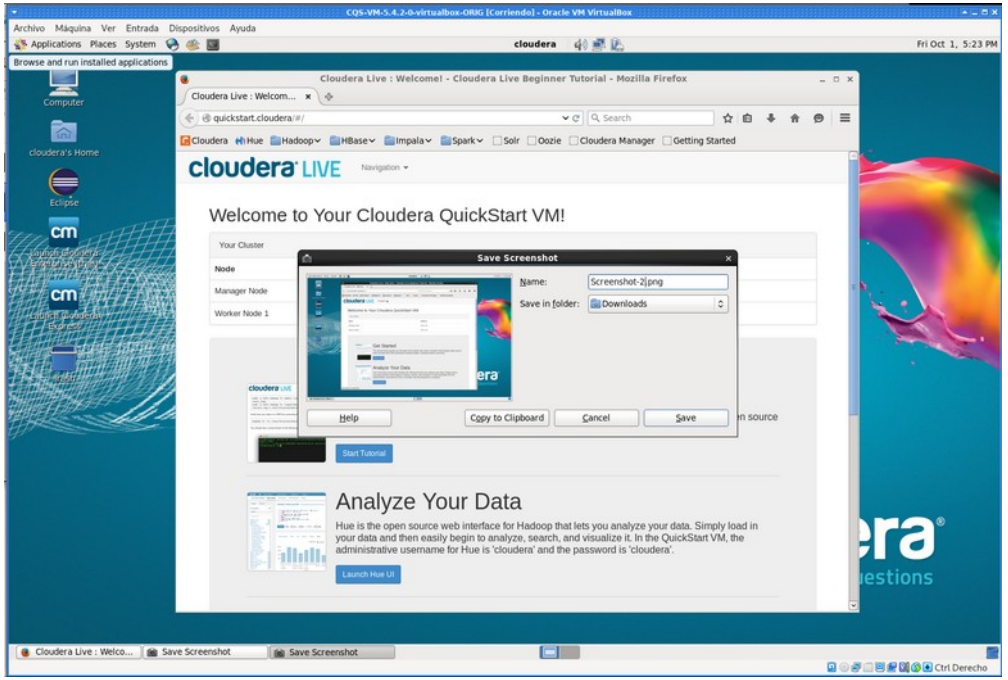
Follow the instructions for installing the Cloudera Virtual Machine:

- Download the extra-configured Cloudera VM: in a Cygwin terminal, execute:

```
date; wget -O cloudera-quickstart-vm-5.4.2-0-virtualbox-GTM.zip
"https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/
EVKUKxkEI0hArIcCBrTbXAEBfc8QNGUDuwBPMIMymJF-HA?e=1WEl4y&download=1";date
```

- Follow the detailed instructions in Moodle
- Start the VM





Configuring the PC for Accessing Directly the LIASOn Clusters

Install the ZeroTier Client and join the virtual private network

Open a terminal window (right-click and launch a terminal as **Administrator**):

zerotier-cli status
zerotier-cli info

zerotier-cli join e5cd7a9e1c7857e5 (liason2)

After finishing, copy the file `hosts` to:
C:/Windows/system32/drivers/etc

notepad C:/Windows/system32/drivers/etc/hosts
(as **Administrator**)

Distribuciones Linux

- Red Hat Enterprise Linux
- Fedora
- CentOS
- Mandriva
- Debian
- Ubuntu
- Mint
- OpenSUSE

Tipos de licencia de software

Open Source

Permite que tanto el código fuente como los archivos binarios sean modificados y redistribuidos libremente.

Ciertas licencias de código abierto pueden incorporar algunas restricciones, como:

- El requisito de mantener el nombre de los autores y la declaración de derechos de autor en el código

- Permitir la modificación del código sólo para usos personales o la redistribución del software para usos no comerciales.

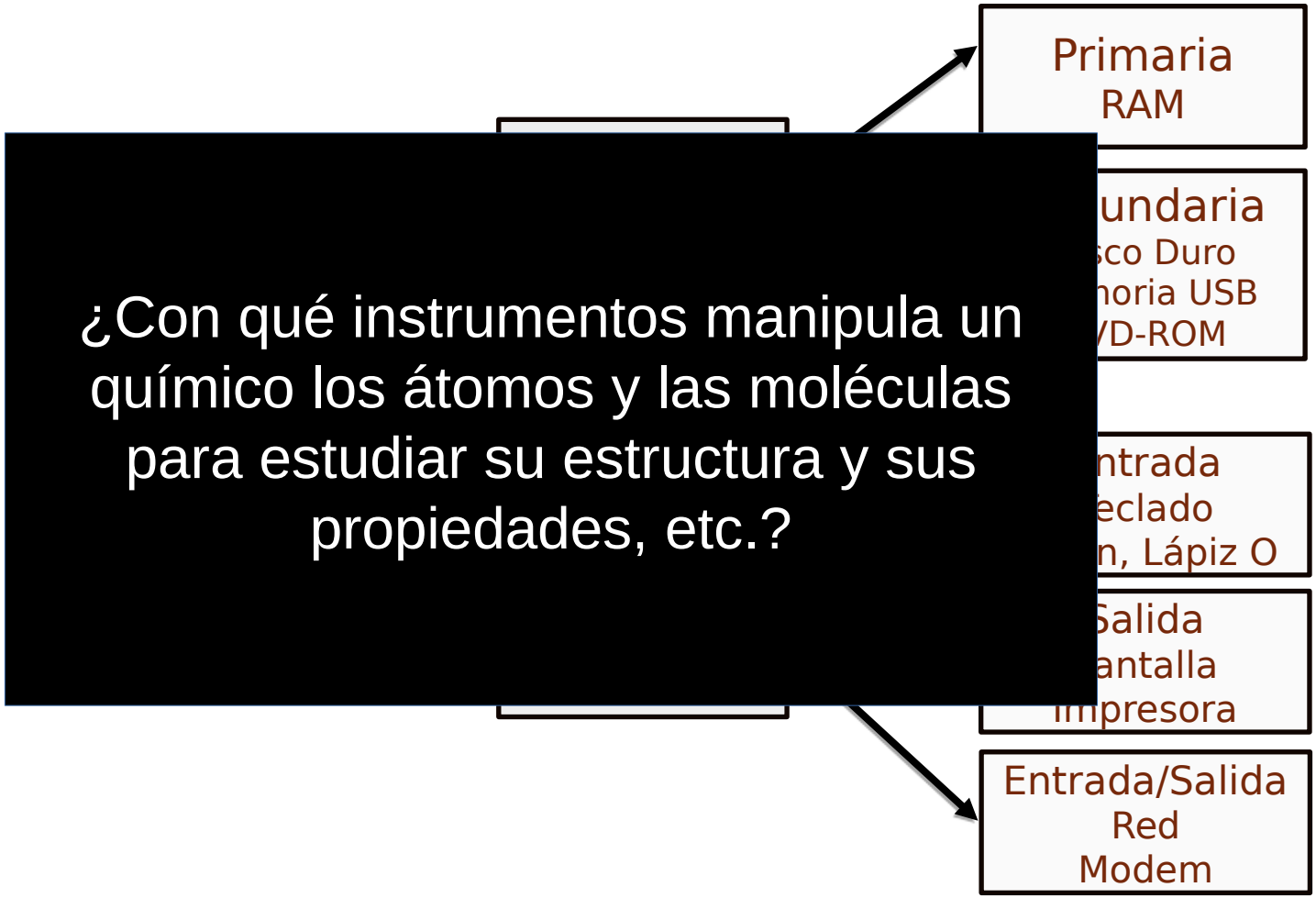
Tipos de licencia de software

- GNU/GPL/LGPL: tal vez la más fiel al espíritu original de Free Software. Nadie puede vender ni comercializar el código ni sus derivados. R es licenciado con GPL v2, aunque no todas sus librerías ¿qué significa eso?

- Berkeley o BSD: permite la libre redistribución, modificación e inclusión en otro software licenciado comercialmente por alguien, para su venta (GPL NO permite esto). PSFL (Python) es compatible con BSD.
- Mozilla Public License o MPL: incluye las cuatro libertades del software libre enunciadas por la Free Software Foundation.
- MIT o X11: Pocas limitaciones en la reutilización y por tanto posee una excelente compatibilidad de licencia.
- Comercial: los derechos se negocian, usualmente pagando un valor en efectivo.

Funciones de un Sistema Operativo

Administrador de los recursos del hardware



¿Con qué instrumentos manipula un químico los átomos y las moléculas para estudiar su estructura y sus propiedades, etc.?

Funciones de un Sistema Operativo

¿Cómo administra los recursos del hardware?

- Multi-usuario
- Multi-procesamiento
- Cada programa se ejecuta de forma concurrente

Monitor de recursos principales del sistema:

- Linux: gnome-system-monitor
- Windows: CTRL + ALT + DEL → select Task Manager then Monitor
  - Poner la ventana en modo “Visible Siempre” y ejecutar en otra ventana el comando top
  - Qué efecto tiene la ejecución de comandos sobre los recursos del sistema?

Downloading Data Files

Open a "Cygwin64 Terminal" and use the commands: df -h to locate a good folder.
To download files from OneDrive: right-click on the file to download (from the web interface) and generate a shareable link, e.g.,
https://icesiedu-my.sharepoint.com/:u:/g/personal/16282252_icesi_edu_co/EXVdFt8IUONHqOWYn5VkFBgBn7kWua6hL-Jqg3y6AWV1uA?e=3o6aNN

Use wget with the following syntax (the quotes are required) including "&download=1" at the end (each of the following in a single line; note that Powerpoint ***incorrectly*** breaks the lines):

```
time wget -O stops.csv
"https://icesiedu-my.sharepoint.com/:x:/g/personal/16282252_icesi_edu_co/EeagjJprX6hMsXsMtKSQYSUBeb43vJa8mx4GIeaLS8hhIQ?e=DuImQN&download=1"

time wget -O stops-shasum.txt
"https://icesiedu-my.sharepoint.com/:t:/g/personal/16282252_icesi_edu_co/Eal2nzrG9PdAjZzFgEjV7lMBzN61kNTQuVB_IgOUoskx6g?e=rs1bYJ&download=1"

time wget -O datagrams.csv
"https://icesiedu-my.sharepoint.com/:x:/g/personal/16282252_icesi_edu_co/EeyEyTkSKg9LoywWHWHAdEgBBqMlAo75Avl346nlTs93vw?e=42EBa0&download=1"

date; wget -O datagrams-shasum.txt
"https://icesiedu-my.sharepoint.com/:t:/g/personal/16282252_icesi_edu_co/EZq8HoyTKgRGiNGNuZI863EBw4z85inh06xWXHMAHwhKMA?e=DyUney&download=1";date
```

Usuarios y Grupos

Linux es multiusuario: varias personas/usuarios pueden trabajar al mismo tiempo en la misma máquina, siendo atendidos por el mismo sistema operativo, desde una misma terminal, o en terminales separadas.

root es el super-usuario (administrador): Linux tiene por configuración un usuario que es el único que tiene todos los permisos y controla el 100% de los recursos del sistema, esto para reducir daños accidentales o maliciosos causados por usuarios inexpertos con excesivos poderes; el superusuario brinda seguridad al O.S.

Los grupos sirven para categorizar y organizar los usuarios. Permite aplicar permisos a grupos que son heredados por los usuarios que formen parte de éstos. Hay grupos de usuarios y grupos de sistema.

Usuarios vs. Archivos y Procesos

Los usuarios tienen propiedad sobre archivos y procesos.

En la máquina virtual de Cloudera o en una CygWin terminal:

- Probar ls –l
- Interpretar la salida del comando anterior
- En Linux: abrir otra ventana de CLI y correr xeyes

Consulta de usuarios y procesos activos:

- w
- top

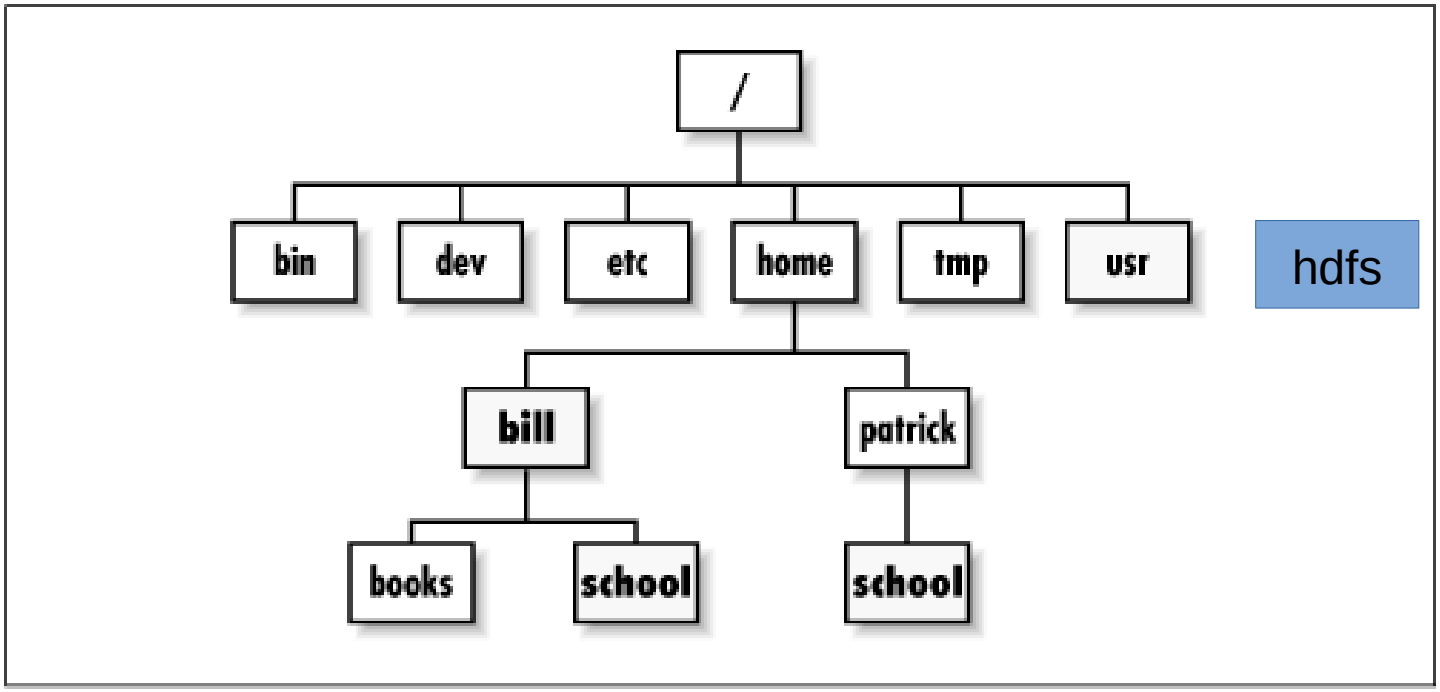
Filesystem

Partición: estructura físico-lógica en que se divide el espacio de almacenamiento de un disco duro.

Filesystem: estructura lógica utilizada para almacenar archivos en una partición.

El sistema de archivos define la forma de almacenar, organizar y recuperar todos los archivos del equipo, para hacer todas las operaciones sobre sus datos eficientemente.

Estructura estándar de directorios



En Linux, las minúsculas son distintas a las mayúsculas (En Windows NO…)

- tree –d –L 1 /
- ls –l
- Verificar usuario propietario y grupo
- Localización de archivos por defecto **(~/Downloads,**
- **~/Documents, ~/Desktop, ~/Pictures, ...)**

File Manager

En Linux, todo archivo y directorio tiene tres niveles de acceso: los que se aplican al propietario del que se aplican al grupo que tiene el archivo y blican a todos los usuarios del sistema. r los permisos cuando listamos un directorio .

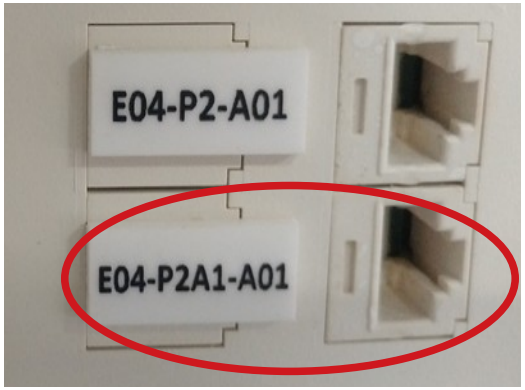
.	Directorio actual
..	Directorio padre
-	Archivo común
d	Directorio
l	Enlace simbólico

rwx	rwx	rwx
usuario	grupo	otros

```
[guested@ice java-example]$ ls -la
total 10316
drwxrwxrwx. 3 guested guested       4096 Oct  9 09:47 .
drwxr-xr-x. 4 guested guested       4096 May 31 10:32 ..
lrwxrwxrwx. 1 guested guested         27 Oct  9 09:47 .bash_history -> /home/guested/.bash_history
drwxrwxr-x. 2 guested vboxusers     4096 Oct  9 09:47 bin
-rw-rw-r--. 1 guested guested   10460713 Feb 18  2019 java.awt.Color
-rw-rw-r--. 1 guested guested    4752142 Feb 18  2019 java.awt.Graphics
-rw-rw-r--. 1 guested guested    5536216 Feb 18  2019 java.util.Scanner
-rwxr-xr-x. 1 guested guested       2599 Feb 19  2019 MyPolygon.class
-rw-rw-r--. 1 guested guested       2695 Feb 19  2019 MyPolygon.java
-rw-rw-r--. 1 guested guested       2101 Feb 19  2019 MyPolygon.java~
-rw-rw-r--. 1 guested guested         45 Feb 18  2019 puntos4.txt
```

Configuración Física de Red

Revisar los sockets RJ45 en la pared para determinar a cuál conectar el cable de red. Por ejemplo:



E04-P2-A01 (nomenclatura del rack institucional) con salida a Internet. **A01** es el número del punto.

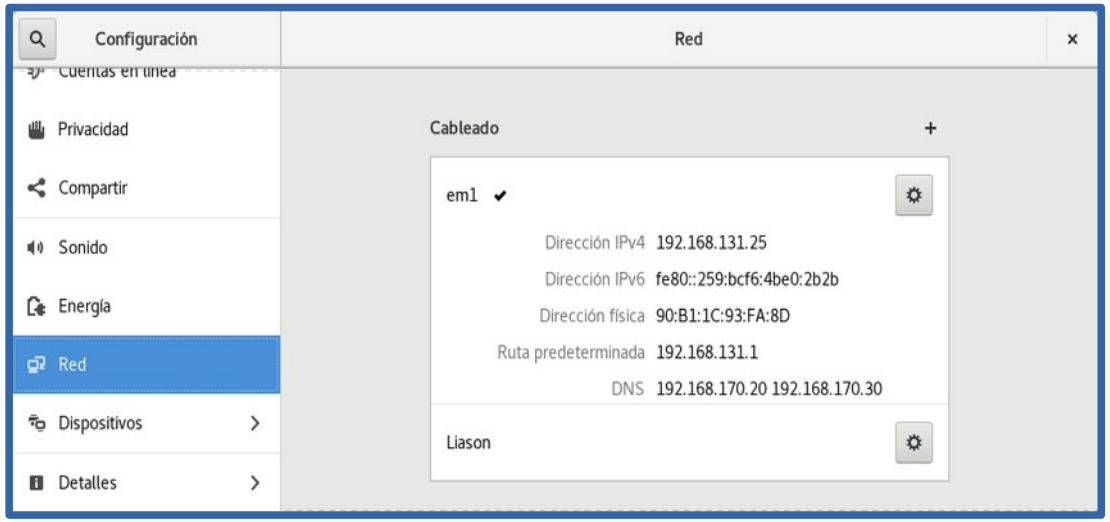
E04-P2A1-A01 (nomenclatura del rack del cluster de los laboratorios LIASON) para usar el cluster y sus servicios internos. **A01** es el número de punto.

Vista en el rack interno (normalmente no deben revisar nada aquí)

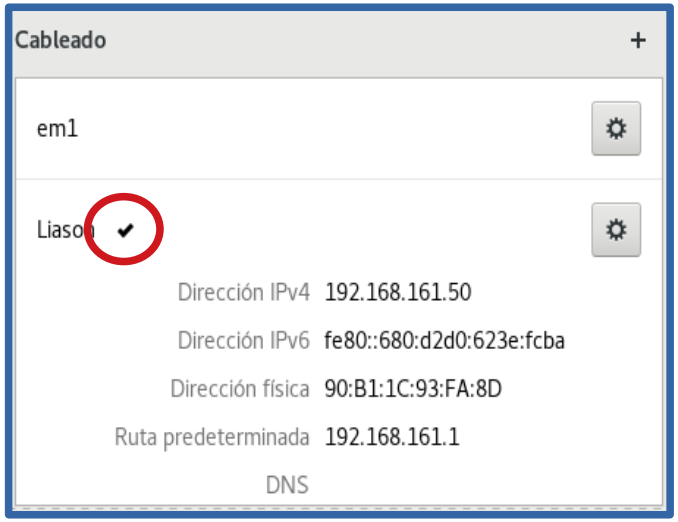
Configuración Lógica de Red

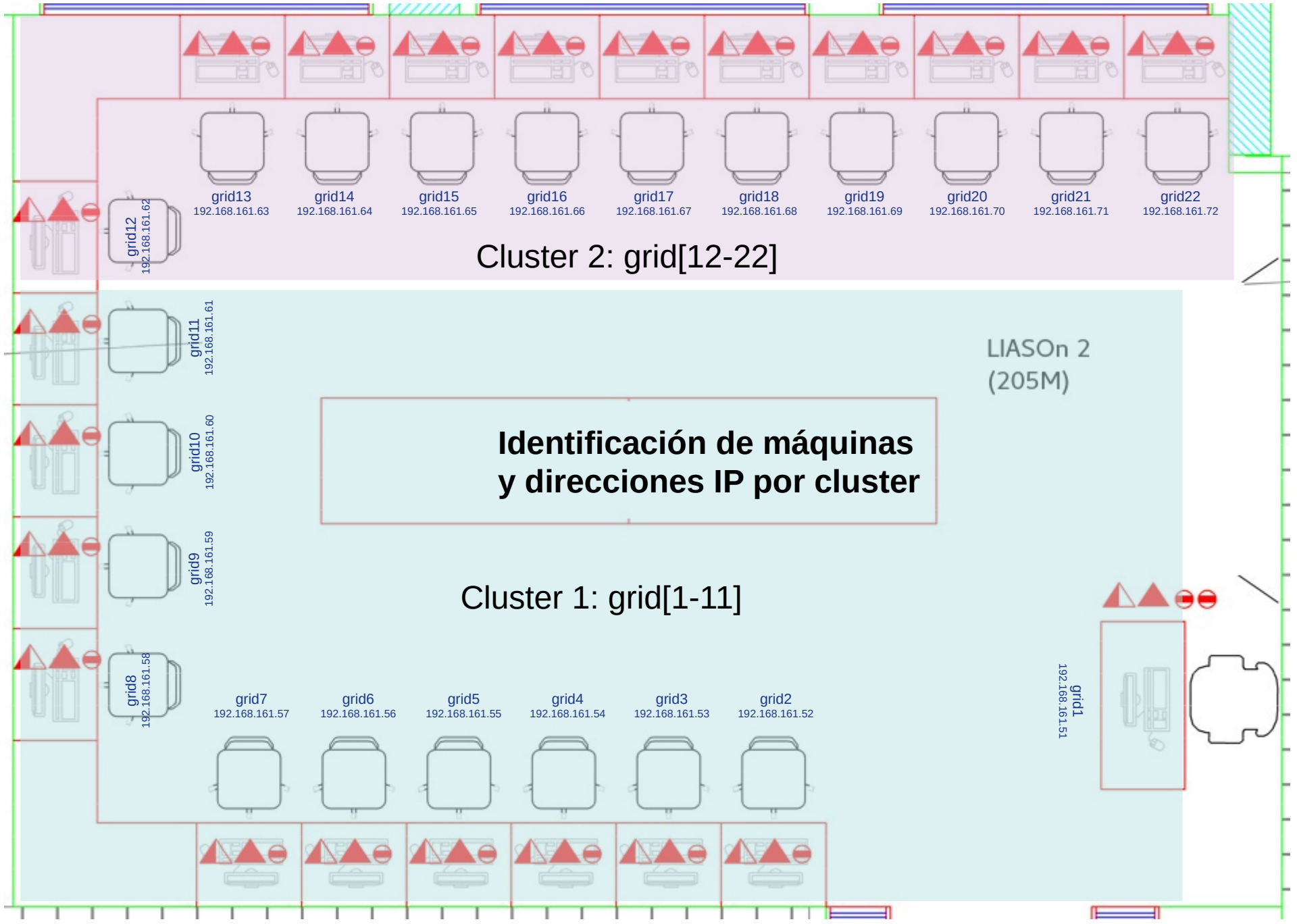
Para funcionar en cluster, además de conectar el cable de red al socket respectivo, debe verificar la configuración lógica (en el sistema operativo del computador).
En la ventana de configuración de red (entrando por el icono en la esquina superior derecha) aparecen dos interfaces:

em1: se utiliza para acceder a internet y a los servicios institucionales de la Universidad Icesi.

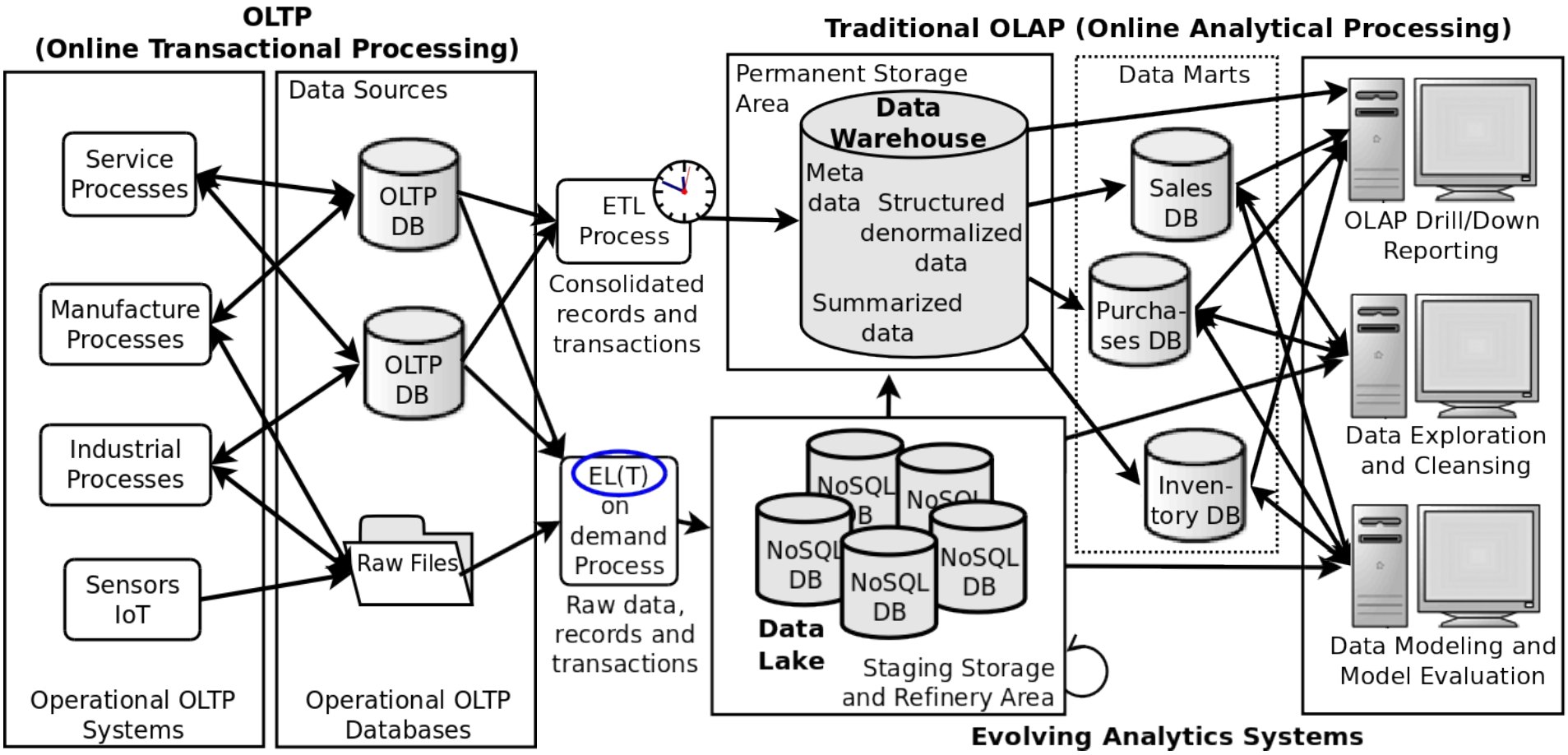


Liason: se utiliza para poner el computador en modo cluster, acceder a la red interna de Liason y a sus servicios internos, por ejemplo de Hadoop y Spark.




grid13
192.168.161.63
grid14
192.168.161.64
grid15
192.168.161.65
grid16
192.168.161.66
grid17
192.168.161.67
grid18
192.168.161.68
grid19
192.168.161.69
grid20
192.168.161.70
grid21
192.168.161.71
grid22
192.168.161.72
grid12
192.168.161.62
Cluster 2: grid[12-22]
grid11
192.168.161.61
grid10
192.168.161.60
grid9
192.168.161.59
grid8
192.168.161.58
LIASOn 2
(205M)
Identificación de máquinas
y direcciones IP por cluster
Cluster 1: grid[1-11]
grid7
192.168.161.57
grid6
192.168.161.56
grid5
192.168.161.55
grid4
192.168.161.54
grid3
192.168.161.53
grid2
192.168.161.52
grid1
192.168.161.51

Big Data and Data Lakes



Configuración Lógica de Red

C:/Windows/
system32/
drivers/etc/hosts

El archivo **/etc/hosts** especifica todas las máquinas de los laboratorios:

Liason1

Direccion IP	Nombre.Dominio	Nombre
192.168.161.21	hgrid1.icesi.edu.co	hgrid1
192.168.161.22	hgrid2.icesi.edu.co	hgrid2
192.168.161.23	hgrid3.icesi.edu.co	hgrid3
192.168.161.24	hgrid4.icesi.edu.co	hgrid4
192.168.161.25	hgrid5.icesi.edu.co	hgrid5
192.168.161.26	hgrid6.icesi.edu.co	hgrid6
192.168.161.27	hgrid7.icesi.edu.co	hgrid7
192.168.161.28	hgrid8.icesi.edu.co	hgrid8
192.168.161.29	hgrid9.icesi.edu.co	hgrid9
192.168.161.30	hgrid10.icesi.edu.co	hgrid10
192.168.161.31	hgrid11.icesi.edu.co	hgrid11
192.168.161.32	hgrid12.icesi.edu.co	hgrid12
192.168.161.33	hgrid13.icesi.edu.co	hgrid13
192.168.161.34	hgrid14.icesi.edu.co	hgrid14
192.168.161.35	hgrid15.icesi.edu.co	hgrid15
192.168.161.36	hgrid16.icesi.edu.co	hgrid16
192.168.161.37	hgrid17.icesi.edu.co	hgrid17
192.168.161.38	hgrid18.icesi.edu.co	hgrid18
192.168.161.9	hgrid19.icesi.edu.co	hgrid19
192.168.161.10	hgrid20.icesi.edu.co	hgrid20
192.168.161.11	hgrid21.icesi.edu.co	hgrid21
192.168.161.12	hgrid22.icesi.edu.co	hgrid22

Liason2

Direccion IP	Nombre.Dominio	Nombre
192.168.161.51	grid1.icesi.edu.co	grid1
192.168.161.52	grid2.icesi.edu.co	grid2
192.168.161.53	grid3.icesi.edu.co	grid3
192.168.161.54	grid4.icesi.edu.co	grid4
192.168.161.55	grid5.icesi.edu.co	grid5
192.168.161.56	grid6.icesi.edu.co	grid6
192.168.161.57	grid7.icesi.edu.co	grid7
192.168.161.58	grid8.icesi.edu.co	grid8
192.168.161.59	grid9.icesi.edu.co	grid9
192.168.161.60	grid10.icesi.edu.co	grid10
192.168.161.61	grid11.icesi.edu.co	grid11
192.168.161.62	grid12.icesi.edu.co	grid12
192.168.161.63	grid13.icesi.edu.co	grid13
192.168.161.64	grid14.icesi.edu.co	grid14
192.168.161.65	grid15.icesi.edu.co	grid15
192.168.161.66	grid16.icesi.edu.co	grid16
192.168.161.67	grid17.icesi.edu.co	grid17
192.168.161.68	grid18.icesi.edu.co	grid18
192.168.161.69	grid19.icesi.edu.co	grid19
192.168.161.70	grid20.icesi.edu.co	grid20
192.168.161.71	grid21.icesi.edu.co	grid21
192.168.161.72	grid22.icesi.edu.co	grid22

Servers Liason

Direccion IP	Nombre.Dominio	Nombre
192.168.161.41	grid100.icesi.edu.co	grid100
192.168.161.42	grid101.icesi.edu.co	grid101
192.168.161.43	grid102.icesi.edu.co	grid102
192.168.161.44	grid103.icesi.edu.co	grid103

UNIVERSIDAD ICESI

Chequeo de Configuración de Red

- Parámetros de Red: el comando ifconfig
  - Interfaz de red (nombre)
  - Dirección IP y hostname
  - Puertos y protocolos de comunicación
- Red cableda
- Red inalámbrica
- ping
- nmap

Emulador de Terminal: Manejo de Archivos

- Buscar archivos (find)
- Copiar localmente (cp)
- Copiar remotamente (scp)
- Mover (mv)
- Borrar (rm)
- El navegador de archivos CLI: mc

Partition and File Commands: Starting to Know About Your Data and How to Manage It

Useful Commands:

- df -h [partition | filePattern]
- du -h [partition | filePattern]
- file inputFilePattern → (text file, binary file) + encoding (raw, base64, zip, gzip, …)
- tree [filePattern]→ display a tree structure of filePattern
- cd [directorio]
- wc -l inputFilePattern → count lines [words, chars] in inputFilePattern
- head -n number inputFilePattern → count lines [words, chars] in inputFilePattern
- tail -n number inputFilePattern → count lines [words, chars] in inputFilePattern

Reconstituting the Datagrams Dataset

- Split files are more easily transferred. In the destination, reconstruct the original file:

1. download the .zip files
2. test the integrity of .zip files:

date; zip -T datag-p1.zip; date
zip -T datag-p2.zip
zip -T datag-p3.zip
(No errors should occur)

3. Decompress files:

date; unzip datag-p1.zip; unzip datag-p2.zip; unzip datag-p3.zip; date

4. Reconstruct the original:

cat datag-p1 datag-p2 datag-p3 > datagrams.zip

5. Test integrity and decompress the original file:

date; zip -T datagrams.zip; date
date; unzip datagrams.zip; date

Understand the Business through Its Data

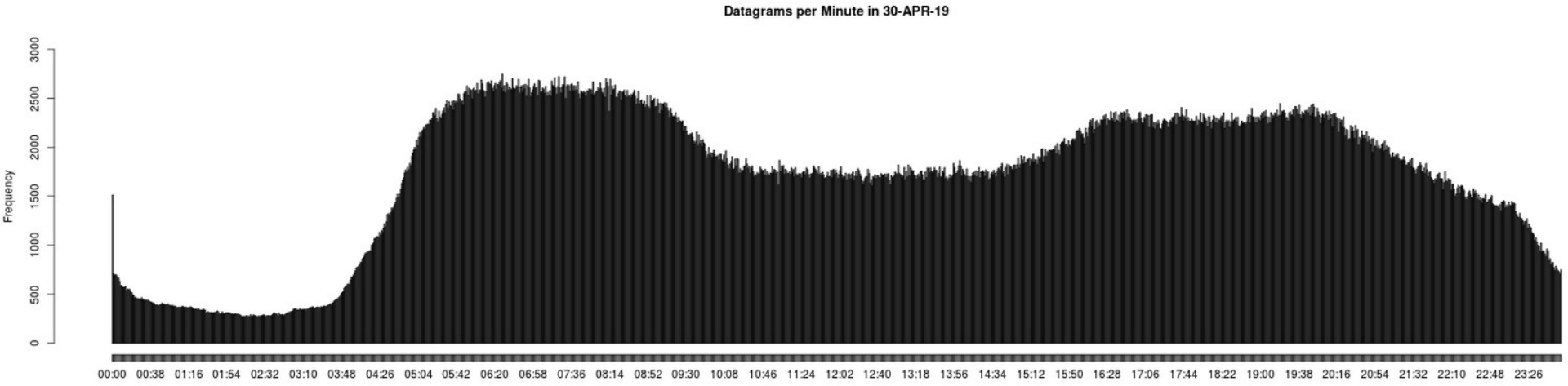
Getting to Know the Data:
How to Start
(and manage)
the Conversation with Large Data?

"More than anything, what data scientists do is to make discoveries while swimming in data.
[…]
data scientists help decision makers shift from ad hoc analysis to an ongoing conversation with data"

Data Scientist: The Sexiest Job of the 21st Century - Harvard Business Review, Oct. 2012

A Short Talk with the Datagrams Dataset

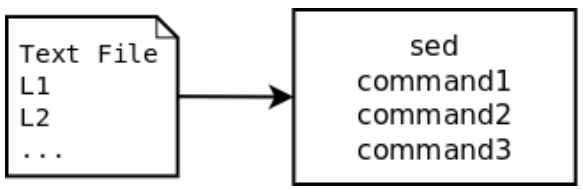
- du -h datagrams.zip; du -h datagrams.csv
- How many rows does datagrams.csv have? How much time did that take?
  Note: this can take more than 10 minutes in a regular PC.
- Show the first and last 10 rows of datagrams.csv
- What is the date format in datagrams.csv?
- Which data does datagrams.csv contain in its columns?
- What does this dataset say to you about the SITM-MIO business?



sed: Stream Editor (far from a perfect tool, but...)

Performs non-interactive and automatic basic text transformations, based on specified commands, on an input stream (a file or input from a pipeline).

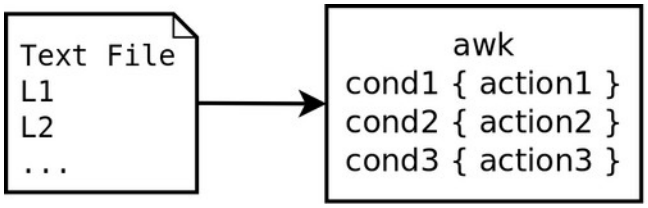
- sed reads the input file sequentially; for each line, it applies each of the specified commands in the specified order. The next command operates on the same line, which may have been modified by earlier commands.
- Commands can be specified in “script” files for replacing, deleting, printing, inserting transforming, lines (i.e., useful for ETL tasks?).

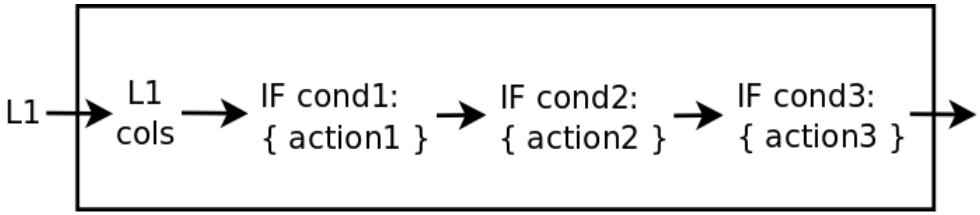


awk: A Domain-Specific Language for ETL Text (Columnar) Processing

Useful to perform non-interactive data extraction, transformation and reporting tasks on an input stream (a file or input from a pipeline).

- Data-driven scripting language consisting of a set of actions to be applied to streams of text data
- Focused on processing strings, associative arrays (that is, arrays indexed by key strings), and regular expressions.
- Reads the input file one line at a time. Each line is split in columns by a char-separator and then tested for a series of conditions in the “program”; if the condition evaluates to true, the corresponding action is executed.





Awk Programs: Patterns and Actions

- An AWK program is a list of condition-action pairs, written as:

```
cond1 { action1 }
cond2 { action2 }
...
```

- Cond is a boolean expression or a /regular-expression/
- Action: a sequence of commands

awk [OPTIONS]… ' program ' [input-file] …

- Useful Options:

-F character used as column separator (precederlo con \ si es un carácter propio decomandos de terminal)

-f scriptCommandsFile

Pseudo variables (field variables): $1, $2, $3, and so on ($0 represents the entire record).

Examples: (download and unzip output-30-APR-19-500k-nonstd.zip from Moodle)

- awk -F\, ' { print $6 " " $5 }' output-30-APR-19-500k-nonstd.csv > nuevoArchivo.csv
- awk -F\, ' { print $6 " " $5 }' output-30-APR-19-500k-nonstd.csv | more
- awk -F\, ' { print $11 }' output-30-APR-19-500k-nonstd.csv

Awk Variables

Built-in variables:

- FILENAME: Contains the name of the current input-file.
- NF: 'N'umber of 'F'ields: contains the number of fields in the current input record. The last field in the input record can be designated by $NF, the 2nd-to-last field by $(NF-1), the 3rd-to-last field by $(NF-2), etc.
- FS: 'F'ield 'S'eparator: the "field separator" character used to split fields in the input record. The default, "white space", includes any space and tab characters. FS can be reassigned to another character to change the field separator.
- NR: 'N'umber of 'R'ecords: the number of input records read so far from all input data files. It starts at zero, but is never automatically reset to zero.
- FNR: 'F'ile 'N'umber of 'R'ecords: the number of input records read so far in the current file. This variable is automatically reset to zero each time a new input file is started.

Examples:

- awk -F\, ' { print NR "," NF ", " $0 }' output-30-APR-19-500k-nonstd.csv > archivo.csv
- awk -F\, ' { print $11 }' output-30-APR-19-500k-nonstd.csv

Awk Conditions

Pseudo Conditions:

- BEGIN : always TRUE before the first line of the input file
- END : always TRUE after the last line of the input file

Examples:

```
NR==1 { print “This is the first line: ” $0 }
NR>1      { print this line appears after 1 }
$2 ~ /reg-exp/ { print $3 }
```

Exercise:

- Download output-30-APR-19-500k-nonstd.csv from Moodle
- How many records does output-30-APR-19-500k-nonstd.csv contain?
- Does all records in output-30-APR-19-500k-nonstd.csv have the same number of columns?

awk: Regular Expressions

Regular expressions (regexp): must be enclosed between slashes: /regexp/

Special symbols:
. : denotes any symbol, one time
* : denotes zero or more times the preceding symbol
+ : denotes one or more times the preceding symbol
? : denotes zero or one time the preceding symbol
(E1 | E2 | …) : denotes the occurrence of E1 or E2 or ...
^ : occurrence at the start of the record
$: occurrence at the end of the record
[a-c] : any of a, b or c

Gabriel
Pedro
Juan
María
[A-Z][a-z]+

awk Structure

```
BEGIN { suma=0; cont=0 }

NR==1   { }

NF>11  {       print cont " : " NF ": $11","$12"
                  suma++
}


END {   print "Filas con cols distintas 12: " suma
}
```

awk: Control Structures

Conditionals:

```
if (condition)
   action-1
else
   Action-2
```

Loops:

```
for (initialization; condition; increment/decrement)
   action


while (condition)
   action
```

awk: Subroutines

```
# Returns maximum number
function find_max(num1, num2){
   if (num1 > num2)
         return num1
   return num2
}

# Main function
function main(num1, num2){
   # Find maximum number
   result = find_max(10, 20)
   print "Maximum =", result
}

# Script execution starts here
BEGIN {
   main(10, 20)
}
```

https://www.gnu.org/software/gawk/manual/html_node/Regexp.html

awk: Exercises

1. Extract from wordsx4.txt all lines between 2019/09/01 and 2019/09/31

2. Extract all lines from wordsx4.txt between 2019/09/05 and 2019/09/07

3. Extract all lines between 2019/09/09 and 2019/09/11 in wordsx4.txt

5. Show the first and last 10 rows of datagrams.csv

6. What is the date format in datagrams.csv?

So far, what data does datagrams.csv contain in its columns?

```
$2 ~ /2019\/09\/[0-3][0-9]/                { print $0                                }
```