

Taller de Crawling

Objetivo

Descargar de forma automática 1000 páginas de Wikipedia.

Procedimiento

1. Abrir una terminal.

- Para abrir una terminal, siga las instrucciones de los ítems número 1 y 2 de la **Guía de conexión**.

2. Uso de wget.

- Para conocer cómo se usa y las opciones o parámetros que tiene esta función se accede a la ayuda escribiendo en la terminal una de las instrucciones presentadas a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ wget -h
```

Bloque de código 1

```
~$ wget --help
```

Bloque de código 2

Al ejecutar esta instrucción en la línea de comandos, obtendremos una ayuda como la que se muestra en la siguiente figura:

```
cursobigdata@cursobigdata:~/Notebooks$ wget -h
GNU Wget 1.15, a non-interactive network retriever.
Usage: wget [OPTION]... [URL]...

Mandatory arguments to long options are mandatory for short options too.

Startup:
  -V, --version             display the version of Wget and exit.
  -h, --help                print this help.
  -b, --background          go to background after startup.
  -e, --execute=COMMAND     execute a '.wgetrc'-style command.

Logging and input file:
  -o, --output-file=FILE    log messages to FILE.
  -a, --append-output=FILE  append messages to FILE.
  -d, --debug               print lots of debugging information.
  -q, --quiet               quiet (no output).
  -v, --verbose              be verbose (this is the default).
  -nv, --no-verbose         turn off verbosity, without being quiet.
  -r, --report-speed=TYPE   Output bandwidth as TYPE. TYPE can be bits.
  -i, --input-file=FILE     download URLs found in local or external FILE.
  -F, --force-html          treat input file as HTML.
  -B, --base=URL            resolves HTML input-file links (-i -F)
                           relative to URL.
  --config=FILE             Specify config file to use.

Download:
  -t, --tries=NUMBER        set number of retries to NUMBER (0 unlimits).
  -r, --retry-connrefused   retry even if connection is refused.
  -O, --output-document=FILE write documents to FILE.
  -nc, --no-clobber         skip downloads that would download to
                           existing files (overwriting them).
  -c, --continue            resume getting a partially-downloaded file.
```

Figura 1

3. Obtener las URLs de las páginas a descargar.

Ingresa a la página principal de Wikipedia (<https://es.wikipedia.org>).



Figura 2

En el panel situado en el costado izquierdo de la página encontrará un enlace nombrado **Página aleatoria**, que carga una página al azar cada vez que se le da clic. La ubicación de este enlace es ilustrada en la figura anterior.

Haga clic derecho sobre el enlace y copie su dirección o ruta. La ruta del enlace es: **<https://es.wikipedia.org/wiki/Especial:Aleatoria>**.

4. Crear una carpeta dentro de su carpeta de trabajo.

Para ver la ruta de la ubicación actual, se escribe en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ pwd
```

Bloque de código 3

La ruta de su ubicación actual deberá ser **/home/cursobigdata/Notebooks**.

Para crear una carpeta que almacena la página web que se va a descargar dentro de su carpeta de trabajo (creada previamente en la Guía de conexión, y que posee la ruta **/home/cursobigdata/Notebooks/CODE**, donde **CODE** corresponde a su **CÓDIGO**), existen dos formas:

- Para crear una carpeta llamada **Taller1-01** desde su carpeta personal, primero debe acceder a esta última (su carpeta personal) y posteriormente crear dicha carpeta. Escriba en la terminal cada uno de los comandos que aparecen a continuación luego del símbolo **\$** y presione la tecla ENTER después de ingresar cada uno de ellos (cada **~\$** indica un comando diferente). Reemplace **CODE** por su **CÓDIGO** para acceder a su carpeta de trabajo, tal como se vio en el ítem número 3 de la Guía previa.

```
~$ cd CODE
~$ mkdir Taller1-01
```

Bloque de código 4

- Para crear la carpeta **Taller1-01** manteniendo su ubicación actual (**/home/cursobigdata/Notebooks/**), ingrese en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER. Reemplace **CODE** por su **CÓDIGO** para acceder a su carpeta de trabajo.

```
~$ mkdir /home/cursobigdata/Notebooks/CODE/Taller1-01
```

Bloque de código 5

Verifique la creación de la carpeta **Taller1-01** dentro de su carpeta de trabajo.

- Si se encuentra dentro de su carpeta de trabajo, escriba en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ ls
```

Bloque de código 6

- De lo contrario, puede verificarlo haciendo uso de la ruta de su carpeta de trabajo. Reemplace **CODE** por su **CÓDIGO** para acceder a su carpeta de trabajo.

```
~$ ls /home/cursobigdata/Notebooks/CODE
```

Bloque de código 7

Listando los archivos y carpetas existentes dentro de su carpeta de trabajo con ayuda de las instrucciones dadas podrá observar que la nueva carpeta ha sido creada, tal como se muestra en la **Figura 3**.

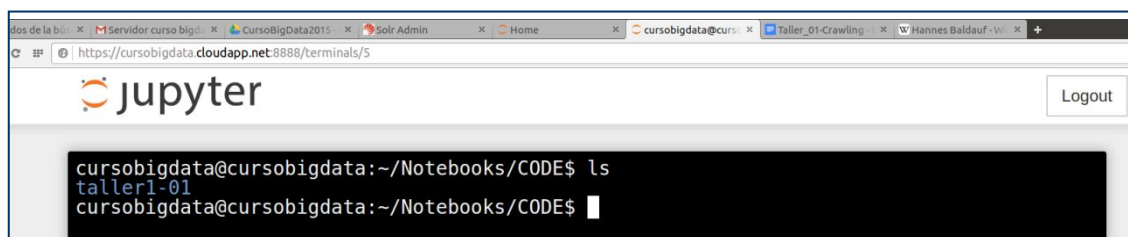


Figura 3

5. Obtener la ruta de la carpeta creada.

Para obtener la ruta de la carpeta creada:

- Acceda a dicha carpeta (**Taller1-01**) desde su carpeta personal escribiendo en la terminal el comando que aparece a continuación luego del símbolo \$ y presione la tecla ENTER.

```
~$ cd Taller1-01
```

Bloque de código 8

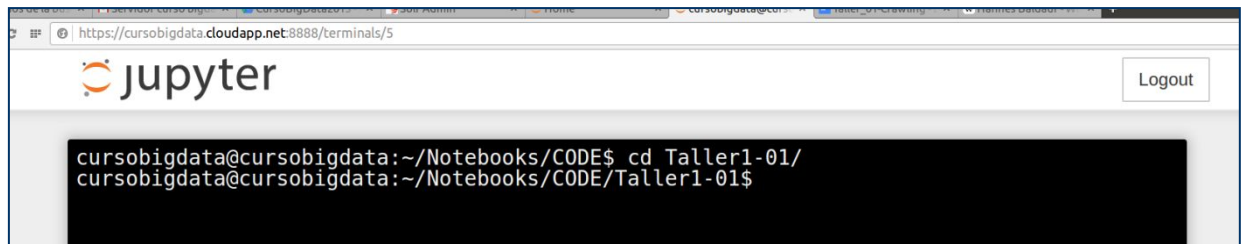


Figura 4

Para ver la ruta de la ubicación actual, es decir, la ruta de la carpeta **Taller1-01**, se escribe en la terminal el comando que aparece a continuación luego del símbolo \$ y presione la tecla ENTER.

```
~$ pwd
```

Bloque de código 9

Su ruta deberá ser **/home/cursobigdata/Notebooks/CODE/Taller1-01** donde **CODE** es el nombre de su carpeta de trabajo y corresponde a su **CÓDIGO**. Tenga en cuenta la ruta de esta carpeta, ya que será empleada a continuación.

6. Descargar una página web de Wikipedia en un directorio local.

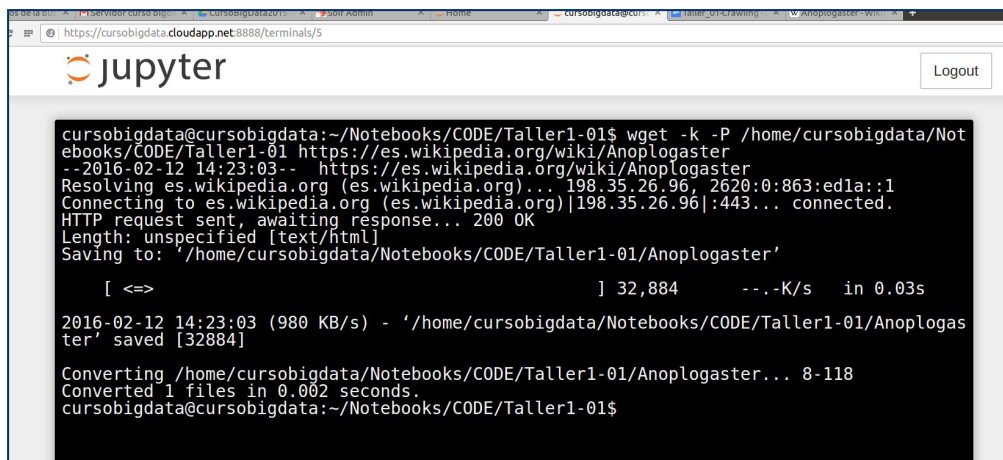
Haciendo uso de **wget**, de la **URL** para ver una página al azar de Wikipedia (obtenida previamente) y la ruta de la carpeta **Taller1-01** se descargará una página web.

En la terminal escriba lo siguiente luego del símbolo \$ y presione la tecla ENTER. Recuerde que **CODE** corresponde a su **CÓDIGO**, reemplácelo.

```
~$ wget -k -P /home/cursobigdata/Notebooks/CODE/Taller1-01  
https://es.wikipedia.org/wiki/Especial:Aleatoria
```

Bloque de código 10

En la **Figura 5** se muestra la descarga de una página web al azar de Wikipedia en el directorio especificado. Cabe resaltar que de no haber sido creada la carpeta **Taller1-01** indicada en la ruta, ésta será creada automáticamente.



```
cursorbigdata@cursorbigdata:~/Notebooks/CODE/Taller1-01$ wget -k -P /home/cursorbigdata/Not  
ebooks/CODE/Taller1-01 https://es.wikipedia.org/wiki/Anoplogaster  
--2016-02-12 14:23:03-- https://es.wikipedia.org/wiki/Anoplogaster  
Resolving es.wikipedia.org (es.wikipedia.org)... 198.35.26.96, 2620:0:863:ed1a::1  
Connecting to es.wikipedia.org (es.wikipedia.org)|198.35.26.96|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: unspecified [text/html]  
Saving to: '/home/cursorbigdata/Notebooks/CODE/Taller1-01/Anoplogaster'  
  
[ <=> ] 32,884 --K/s in 0.03s  
  
2016-02-12 14:23:03 (980 KB/s) - '/home/cursorbigdata/Notebooks/CODE/Taller1-01/Anoplogas  
ter' saved [32884]  
  
Converting /home/cursorbigdata/Notebooks/CODE/Taller1-01/Anoplogaster... 8-118  
Converted 1 files in 0.002 seconds.  
cursorbigdata@cursorbigdata:~/Notebooks/CODE/Taller1-01$
```

Figura 5

7. Verificar la descarga de una página web de Wikipedia.

Si la ruta de su ubicación actual es **/home/cursobigdata/Notebooks/CODE/Taller1-01** dónde **CODE** corresponde a su **CÓDIGO**, es decir, dentro de la carpeta **Taller1-01**, escriba en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER. Compare la ruta obtenida.

```
~$ pwd
```

Bloque de código 11

Ya que se encuentra dentro de la carpeta **Taller1-01**, liste los archivos existentes escribiendo en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ ls
```

Bloque de código 12

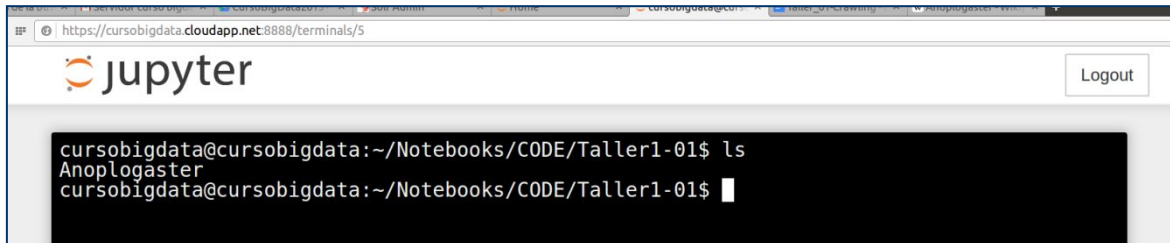


Figura 6

De lo contrario, puede verificarlo haciendo uso de la ruta **/home/cursobigdata/Notebooks/CODE/Taller1-01**. Reemplace **CODE** por su **CÓDIGO**.

```
~$ ls /home/cursobigdata/Notebooks/CODE/taller1-01
```

Bloque de código 13

8. Descarga automatizada de 1000 páginas web aleatorias de Wikipedia.

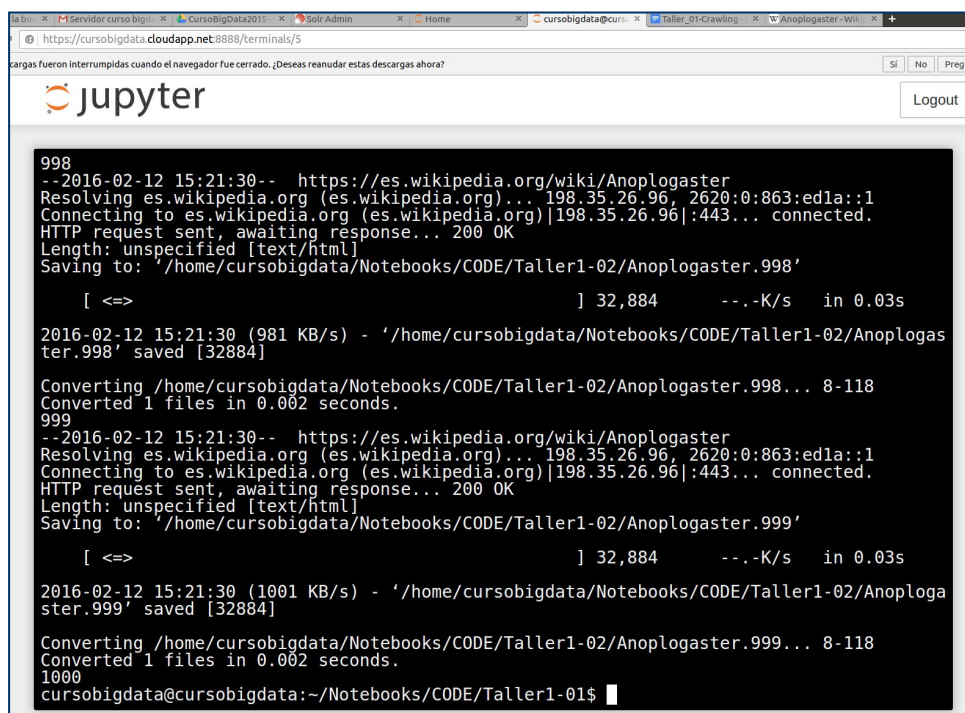
- Desde la shell se crea un script para automatizar el proceso de descarga, descargando **1000** páginas web al azar de Wikipedia, en una carpeta llamada **Taller1-02** que no ha sido creada previamente, pero estará localizada en su carpeta personal. Por lo tanto, la ruta de la carpeta en la que se almacenarán las páginas web descargadas es: **/home/cursobigdata/Notebooks/CODE/Taller1-02** (dónde **CODE** corresponde a su **CÓDIGO**).

Escriba en la terminal los comandos que aparecen a continuación. Digite la instrucción dada luego del símbolo **\$** o **>** y presione la tecla ENTER. Cada símbolo indica un comando y una línea diferente.


```
~$ for i in {1..1000}; do
> wget -k -P /home/cursobigdata/Notebooks/CODE/Taller1-02
https://es.wikipedia.org/wiki/Anoplogaster;
> echo $i;
> done
```

Bloque de código 14

Una vez digitada la ultima linea del codigo, luego de presionar la tecla ENTER la descarga de las páginas web inicia, y termina luego de un par de minutos.



```
998
--2016-02-12 15:21:30-- https://es.wikipedia.org/wiki/Anoplogaster
Resolving es.wikipedia.org (es.wikipedia.org)... 198.35.26.96, 2620:0:863:ed1a::1
Connecting to es.wikipedia.org (es.wikipedia.org)|198.35.26.96|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: '/home/cursobigdata/Notebooks/CODE/Taller1-02/Anoplogaster.998'

[ <=> ] 32,884 --.-K/s in 0.03s

2016-02-12 15:21:30 (981 KB/s) - '/home/cursobigdata/Notebooks/CODE/Taller1-02/Anoplogas
ter.998' saved [32884]

Converting /home/cursobigdata/Notebooks/CODE/Taller1-02/Anoplogaster.998... 8-118
Converted 1 files in 0.002 seconds.
999
--2016-02-12 15:21:30-- https://es.wikipedia.org/wiki/Anoplogaster
Resolving es.wikipedia.org (es.wikipedia.org)... 198.35.26.96, 2620:0:863:ed1a::1
Connecting to es.wikipedia.org (es.wikipedia.org)|198.35.26.96|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: '/home/cursobigdata/Notebooks/CODE/Taller1-02/Anoplogaster.999'

[ <=> ] 32,884 --.-K/s in 0.03s

2016-02-12 15:21:30 (1001 KB/s) - '/home/cursobigdata/Notebooks/CODE/Taller1-02/Anoploga
ster.999' saved [32884]

Converting /home/cursobigdata/Notebooks/CODE/Taller1-02/Anoplogaster.999... 8-118
Converted 1 files in 0.002 seconds.
1000
cursobigdata@cursobigdata:~/Notebooks/CODE/Taller1-01$
```

Figura 7

9. Visualización de las páginas web descargadas.

Si la ruta de su ubicación actual es **/home/cursobigdata/Notebooks/CODE/Taller1-01** dónde **CODE** corresponde a su **CÓDIGO** debe regresar a su carpeta personal y acceder a la carpeta **Taller1-02**. Escriba en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER, compare la ruta obtenida.


```
~$ pwd
```

Bloque de código 15

Para regresar a su carpeta personal y acceder a la carpeta **Taller1-02**, escriba en la terminal cada uno de los comandos que aparecen a continuación luego del símbolo **\$** y presione la tecla ENTER después de ingresar cada uno de ellos. Cada **~\$** indica un comando diferente.

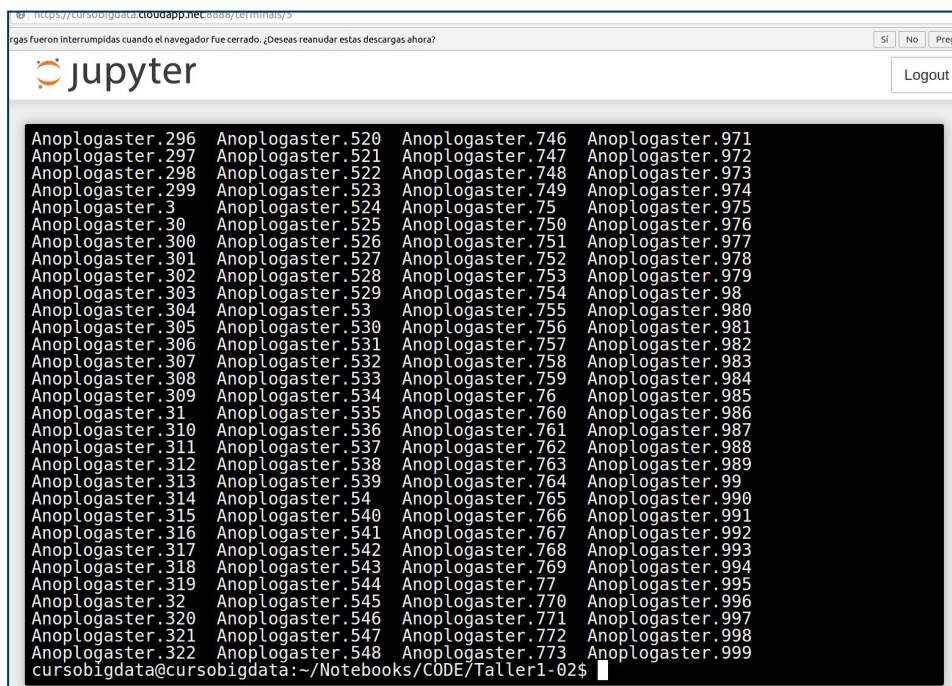
```
~$ cd ..  
~$ cd Taller1-02
```

Bloque de código 16

Una vez dentro de la carpeta señalada, se listan los archivos existentes escribiendo en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ ls
```

Bloque de código 17



The screenshot shows a Jupyter Notebook interface with a terminal window open. The terminal displays a list of files in a directory, organized in four columns. The files are named 'Anoplogaster.' followed by a number. The numbers range from 296 to 999. The terminal prompt at the bottom is 'cursobigdata@cursobigdata:~/Notebooks/CODE/Taller1-02\$'.

```
Anoplogaster.296 Anoplogaster.520 Anoplogaster.746 Anoplogaster.971  
Anoplogaster.297 Anoplogaster.521 Anoplogaster.747 Anoplogaster.972  
Anoplogaster.298 Anoplogaster.522 Anoplogaster.748 Anoplogaster.973  
Anoplogaster.299 Anoplogaster.523 Anoplogaster.749 Anoplogaster.974  
Anoplogaster.3 Anoplogaster.524 Anoplogaster.75 Anoplogaster.975  
Anoplogaster.30 Anoplogaster.525 Anoplogaster.750 Anoplogaster.976  
Anoplogaster.300 Anoplogaster.526 Anoplogaster.751 Anoplogaster.977  
Anoplogaster.301 Anoplogaster.527 Anoplogaster.752 Anoplogaster.978  
Anoplogaster.302 Anoplogaster.528 Anoplogaster.753 Anoplogaster.979  
Anoplogaster.303 Anoplogaster.529 Anoplogaster.754 Anoplogaster.98  
Anoplogaster.304 Anoplogaster.53 Anoplogaster.755 Anoplogaster.980  
Anoplogaster.305 Anoplogaster.530 Anoplogaster.756 Anoplogaster.981  
Anoplogaster.306 Anoplogaster.531 Anoplogaster.757 Anoplogaster.982  
Anoplogaster.307 Anoplogaster.532 Anoplogaster.758 Anoplogaster.983  
Anoplogaster.308 Anoplogaster.533 Anoplogaster.759 Anoplogaster.984  
Anoplogaster.309 Anoplogaster.534 Anoplogaster.76 Anoplogaster.985  
Anoplogaster.31 Anoplogaster.535 Anoplogaster.760 Anoplogaster.986  
Anoplogaster.310 Anoplogaster.536 Anoplogaster.761 Anoplogaster.987  
Anoplogaster.311 Anoplogaster.537 Anoplogaster.762 Anoplogaster.988  
Anoplogaster.312 Anoplogaster.538 Anoplogaster.763 Anoplogaster.989  
Anoplogaster.313 Anoplogaster.539 Anoplogaster.764 Anoplogaster.99  
Anoplogaster.314 Anoplogaster.54 Anoplogaster.765 Anoplogaster.990  
Anoplogaster.315 Anoplogaster.540 Anoplogaster.766 Anoplogaster.991  
Anoplogaster.316 Anoplogaster.541 Anoplogaster.767 Anoplogaster.992  
Anoplogaster.317 Anoplogaster.542 Anoplogaster.768 Anoplogaster.993  
Anoplogaster.318 Anoplogaster.543 Anoplogaster.769 Anoplogaster.994  
Anoplogaster.319 Anoplogaster.544 Anoplogaster.77 Anoplogaster.995  
Anoplogaster.32 Anoplogaster.545 Anoplogaster.770 Anoplogaster.996  
Anoplogaster.320 Anoplogaster.546 Anoplogaster.771 Anoplogaster.997  
Anoplogaster.321 Anoplogaster.547 Anoplogaster.772 Anoplogaster.998  
Anoplogaster.322 Anoplogaster.548 Anoplogaster.773 Anoplogaster.999  
cursobigdata@cursobigdata:~/Notebooks/CODE/Taller1-02$
```

Figura 8

- De lo contrario, puede verificarlo haciendo uso de la ruta **/home/cursobigdata/Notebooks/CODE/Taller1-02**. Reemplace **CODE** por su **CÓDIGO**.

```
~$ ls ls /home/cursobigdata/Notebooks/CODE/Taller1-02/
```

Bloque de código 18

10. Comprobar la existencia de 1000 archivos dentro de la carpeta Taller1-02.

- Si la ruta de su ubicación actual es **/home/cursobigdata/Notebooks/CODE/Taller1-02** dónde **CODE** corresponde a su **CÓDIGO**, es decir, dentro de la carpeta **Taller1-02**, escriba en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER. Compare la ruta obtenida.

```
~$ pwd
```

Bloque de código 19

Una vez dentro de la carpeta señalada, se listan los archivos existentes escribiendo en la terminal el comando que aparece a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ ls | wc -l
```

Bloque de código 20

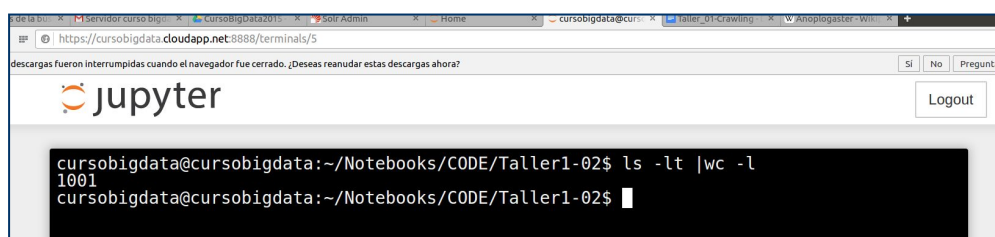


Figura 9

Como se puede observar hay 1000 archivos en la carpeta **Taller1-02**, correspondientes a las páginas web descargadas.

De lo contrario, puede verificarlo haciendo uso de la ruta **/home/cursobigdata/Notebooks/CODE/Taller1-02**. Reemplace **CODE** por su **CÓDIGO**.

```
~$ ls /home/cursobigdata/Notebooks/CODE/ Taller1-02 | wc -l
```

Bloque de código 21

11. Visualizar uno de los archivos descargados.

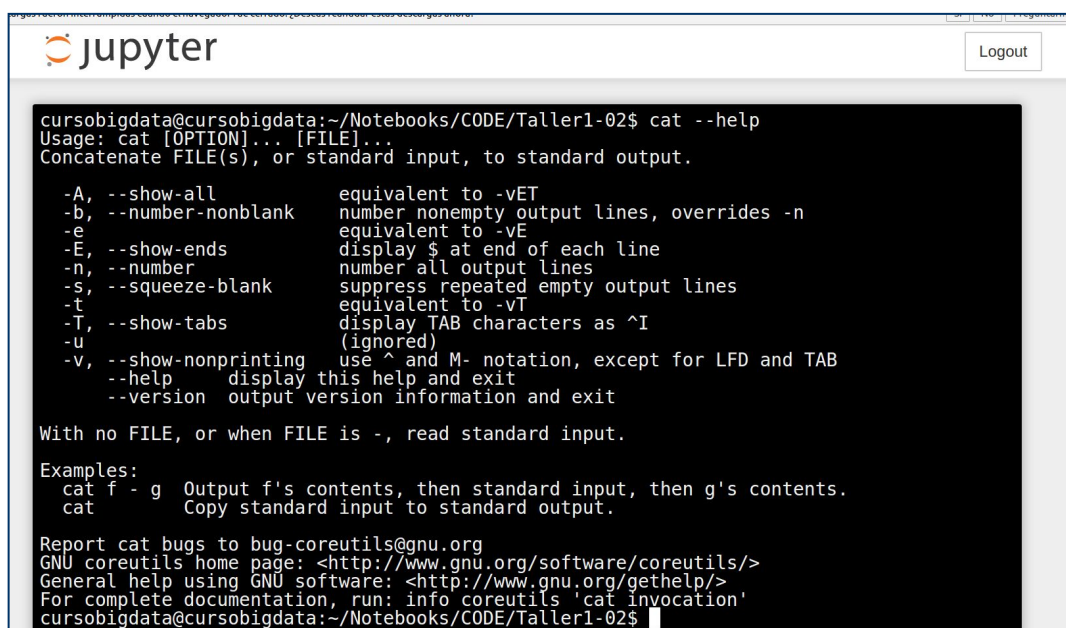
Entre los comandos existentes para visualizar el contenido de un archivo, el comando **cat** permite visualizar el todo el documento desde la terminal desplazándose a lo largo de este con ayuda de la barra de desplazamiento o el scroll del mouse.

Para conocer cómo se usa y las opciones o parámetros que tiene esta función se accede a la ayuda escribiendo en la terminal la instrucción presentada a continuación luego del símbolo **\$** y presione la tecla ENTER.

```
~$ cat --help
```

Bloque de código 22

Al ejecutar esta instrucción en la línea de comandos, obtendremos una ayuda como la que se muestra en la siguiente figura:



```
jupyter
Logout

cursobigdata@cursobigdata:~/Notebooks/CODE/Taller1-02$ cat --help
Usage: cat [OPTION]... [FILE]...
Concatenate FILE(s), or standard input, to standard output.

-A, --show-all           equivalent to -vET
-b, --number-nonblank     number nonempty output lines, overrides -n
-e                        equivalent to -vE
-E, --show-ends           display $ at end of each line
-n, --number              number all output lines
-s, --squeeze-blank       suppress repeated empty output lines
-t                        equivalent to -vT
-T, --show-tabs           display TAB characters as ^I
-u                        (ignored)
-v, --show-nonprinting    use ^ and M- notation, except for LFD and TAB
--help                   display this help and exit
--version                 output version information and exit

With no FILE, or when FILE is -, read standard input.

Examples:
cat f - g   Output f's contents, then standard input, then g's contents.
cat        Copy standard input to standard output.

Report cat bugs to bug-coreutils@gnu.org
GNU coreutils home page: <http://www.gnu.org/software/coreutils/>
General help using GNU software: <http://www.gnu.org/gethelp/>
For complete documentation, run: info coreutils 'cat invocation'
cursobigdata@cursobigdata:~/Notebooks/CODE/Taller1-02$
```

Figura 10

Visualice el contenido del archivo de su preferencia situado en su ubicación actual

(**/home/cursobigdata/Notebooks/CODE/Taller1-02**)

escribiendo en la terminal la instrucción presentada a continuación luego del símbolo \$ y presione la tecla ENTER.

```
~$ cat FILE_NAME
```

Bloque de código 23

Reemplace FILE_NAME por el nombre de uno de los archivos descargados previamente listados con ayuda del comando presente en el **Bloque de código 17** (ver **Figura 8**).

12. Cierre la terminal y desconéctese del servidor.

- Para desconectarse del servidor, siga las instrucciones de los ítems número 5 y 6 de la **Guía de conexión**.

Fin de la Guia