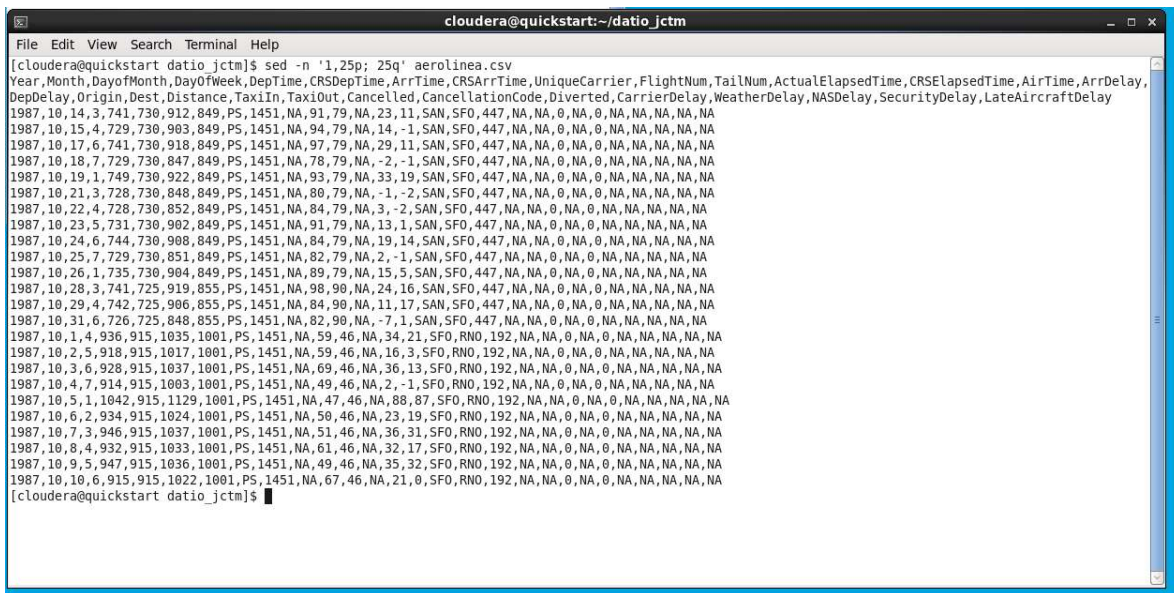


SECCION 1. GNU/LINUX

1.- Del archivo **aerolineas.csv** (el archivo descomprimido que todavía debería estar en su local y no el del HDFS) use comandos de GNU/Linux para obtener las 25 primeras líneas (incluyendo encabezado) **SIN** usar el comando head.

Respuesta : `sed -n '1,25p; 25q' aerolinea.csv`



```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ sed -n '1,25p; 25q' aerolinea.csv
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,
DepDelay,Origin,Dest,Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
1987,10,14,3,741,730,912,849,PS,1451,NA,91,79,NA,23,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,15,4,729,730,903,849,PS,1451,NA,94,79,NA,14,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,17,6,741,730,918,849,PS,1451,NA,97,79,NA,29,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,18,7,729,730,847,849,PS,1451,NA,78,79,NA,-2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,19,1,749,730,922,849,PS,1451,NA,93,79,NA,33,19,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,21,3,728,730,848,849,PS,1451,NA,80,79,NA,-1,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,24,6,744,730,908,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,28,3,741,725,919,855,PS,1451,NA,98,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,5,1,1042,915,1129,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,6,2,934,915,1024,1001,PS,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,7,3,946,915,1037,1001,PS,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,8,4,932,915,1033,1001,PS,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,9,5,947,915,1036,1001,PS,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,10,6,915,915,1022,1001,PS,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
[cloudera@quickstart datio_jctm]$
```

2.-Como ya se ha visto, utilizar el redireccionamiento destructivo (>) implica almacenar típicamente algún contenido en un archivo (ej. **echo "contenido" > archivo**).

Pero lo cierto es que con este comando no se apreciará en pantalla lo que se desea almacenar en dicho archivo, por ello es que se necesita que, con base en el comando resultado del ejercicio 1 y con la investigación del comando **tee**, por un lado el contenido se introduzca en el archivo **ejercicio_2.txt** y por el otro se muestre en pantalla la operación.

Caber mencionar que todo se debe registrar como una sola instrucción, es decir, no se puede ejecutar el resultado por partes, para ello tal vez quiera leer esta liga:

<http://www.linfo.org/pipes.html>

Respuesta : `sed -n '1,25p; 25q' aerolinea.csv | tee ejercicio_2.txt`

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ sed -n '1,25p; 25q' aerolinea.csv | tee ejercicio_2.txt
Year,Month,DayOfMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,
DepDelay,Origin,Dest,Distance,TaxiIn,TaxiOut,Canceled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
1987,10,14,3,741,730,912,849,PS,1451,NA,91,79,NA,23,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,15,4,729,730,903,849,PS,1451,NA,94,79,NA,14,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,17,6,741,730,918,849,PS,1451,NA,97,79,NA,29,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,18,7,729,730,847,849,PS,1451,NA,78,79,NA,-2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,19,1,749,730,922,849,PS,1451,NA,93,79,NA,33,19,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,21,3,728,730,848,849,PS,1451,NA,80,79,NA,-1,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,24,6,744,730,908,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,28,3,741,725,919,855,PS,1451,NA,98,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,5,1,1042,915,1129,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,6,2,934,915,1024,1001,PS,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,7,3,946,915,1037,1001,PS,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,8,4,932,915,1033,1001,PS,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,9,5,947,915,1036,1001,PS,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,10,6,915,915,1022,1001,PS,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
[cloudera@quickstart datio_jctm]$ ls
aerolinea.csv aerolinea.head10.csv ejercicio_2.txt
[cloudera@quickstart datio_jctm]$
```

3.- Cambie el nombre del archivo **ejercicio_2.txt** a **ejercicio_3.txt** SIN usar el comando rename

Respuesta : `cat ejercicio_2.txt > ejercicio_3.txt`

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ cat ejercicio_2.txt > ejercicio_3.txt
[cloudera@quickstart datio_jctm]$ ls
aerolinea.csv aerolinea.head10.csv ejercicio_2.txt ejercicio_3.txt
[cloudera@quickstart datio_jctm]$
```

4.- Con algún comando en GNU/Linux tome las 25 últimas líneas del archivo **aerolínea.csv** SIN emplear el comando tail y guárdelo como **ejercicio_4.txt**


Resultado: `sed -n '123534947,123534972p; 123534972q' aerolinea.csv > ejercicio_4.txt`

A terminal window titled 'cloudera@quickstart: ~/datio_jctm' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
[cloudera@quickstart datio_jctm]$ wc -l aerolinea.csv
123534972 aerolinea.csv
[cloudera@quickstart datio_jctm]$
[cloudera@quickstart datio_jctm]$
[cloudera@quickstart datio_jctm]$ sed -n '123534947,123534972p; 123534972q' aerolinea.csv > ejercicio_4.txt
[cloudera@quickstart datio_jctm]$ r
```

5.- Concatene los archivos **ejercicio_3.txt** y **ejercicio_4.txt** en un archivo **ejercicio_5.txt** y en esa misma pantalla resultado muestre el contenido de **ejercicio_5.txt**

Resultado: `cat ejercicio_3.txt ejercicio_4.txt > ejercicio_5.txt`

A terminal window titled 'cloudera@quickstart: ~/datio_jctm' with a menu bar (File, Edit, View, Search, Terminal, Help). The terminal shows the following commands and output:

```
^[[A[cloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > tee ejercicio_5.txt
cat: ejercicio_5.txt: No such file or directory
[cloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > ejercicio_5.txt
[cloudera@quickstart datio_jctm]$
```

6.- Usando el comando **ls** y sus opciones, verifique el peso de **ejercicio_5.txt**, señalando en la captura de pantalla dónde se encuentra éste.

Resultado: `ls -lh`

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
^[[Acloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > tee ejercicio_5.txt
cat: ejercicio_5.txt: No such file or directory
[cloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > ejercicio_5.txt
[cloudera@quickstart datio_jctm]$ ls -lh
total 12G
-rwxrwxrwx 1 cloudera cloudera 12G Jun 12 16:22 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 14 15:18 aerolinea.head10.csv
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:49 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:53 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 09:20 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:28 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:26 ejercicio_5.txtreset
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:27 tee
[cloudera@quickstart datio_jctm]$
```

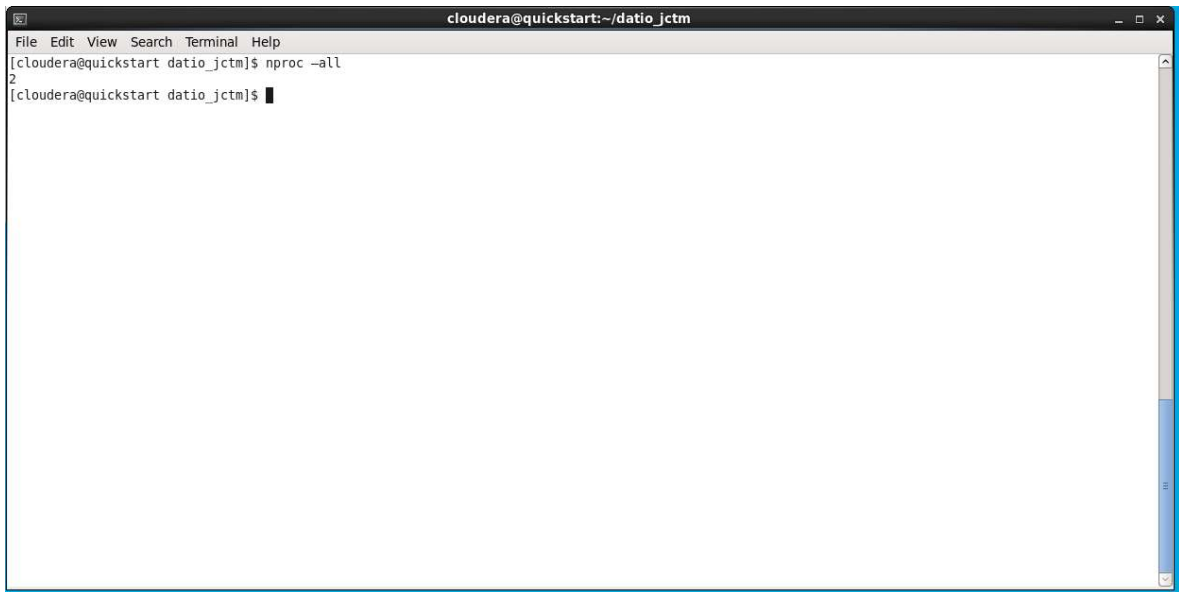
7.- Modifique la fecha de acceso de **ejercicio_5.txt** al 25 de Agosto del 2018 y muestre en pantalla dónde se puede apreciar ese resultado.

Resultado: touch -d "25 Aug 2018" ejercicio_5.txt

```
^[[Acloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > tee ejercicio_5.txt
cat: ejercicio_5.txt: No such file or directory
[cloudera@quickstart datio_jctm]$ cat ejercicio_3.txt ejercicio_4.txt > ejercicio_5.txt
[cloudera@quickstart datio_jctm]$ ls -lh
total 12G
-rwxrwxrwx 1 cloudera cloudera 12G Jun 12 16:22 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 14 15:18 aerolinea.head10.csv
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:49 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:53 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 09:20 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:28 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:26 ejercicio_5.txtreset
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:27 tee
[cloudera@quickstart datio_jctm]$ touch -d "25 Agu 2018" ejercicio_5.txt
touch: invalid date format '25 Agu 2018'
[cloudera@quickstart datio_jctm]$ touch -d "25 Aug 2018" ejercicio_5.txt
[cloudera@quickstart datio_jctm]$ ls -lh
total 12G
-rwxrwxrwx 1 cloudera cloudera 12G Jun 12 16:22 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 14 15:18 aerolinea.head10.csv
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:49 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:53 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 09:20 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Aug 25 2018 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:26 ejercicio_5.txtreset
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:27 tee
[cloudera@quickstart datio_jctm]$
```

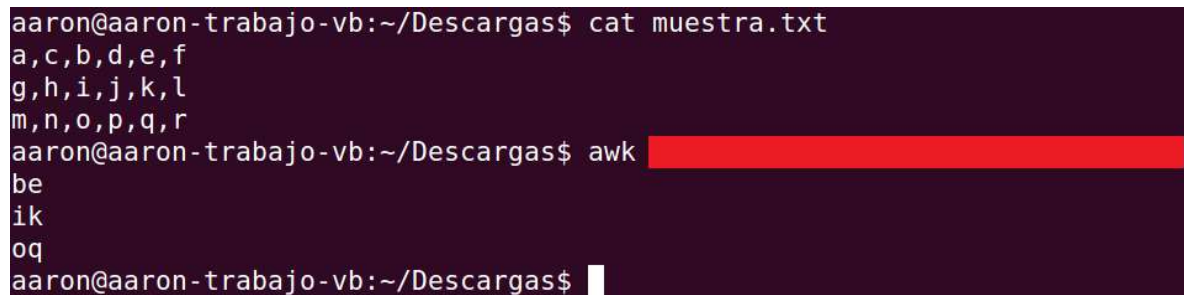
8.- ¿Con cuál comando se puede averiguar el número de núcleos en un sistema GNU/Linux? Investigue y coloque el resultado, haciendo énfasis en el lugar donde se puede apreciar esa información.

Resultado: nproc --all

A terminal window titled "cloudera@quickstart: ~/datio_jctm". The window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The terminal shows the command "[cloudera@quickstart datio_jctm]\$ nproc --all" followed by the output "2". The prompt "[cloudera@quickstart datio_jctm]\$ " is visible at the bottom.

```
cloudera@quickstart: ~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ nproc --all
2
[cloudera@quickstart datio_jctm]$
```

9.- Investigue en qué consiste awk y por medio de esa herramienta imprima en pantalla sólo la tercera y quinta columnas (de izquierda a derecha) del archivo **ejercicio_5.txt**. He aquí un ejemplo de cómo se ve el resultado con otro archivo que no tiene que ver con el curso:

A terminal window with a dark background. It shows the command "cat muestra.txt" and its output: "a,c,b,d,e,f", "g,h,i,j,k,l", and "m,n,o,p,q,r". Then, the command "awk" is entered, followed by a redacted line. The output of the awk command is "be", "ik", and "oq".

```
aaron@aaron-trabajo-vb:~/Descargas$ cat muestra.txt
a,c,b,d,e,f
g,h,i,j,k,l
m,n,o,p,q,r
aaron@aaron-trabajo-vb:~/Descargas$ awk 
be
ik
oq
aaron@aaron-trabajo-vb:~/Descargas$
```

Resultado: `awk -F ',' '{print $3,$5}' ejercicio_5.txt`

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ awk -F ',' '{print $3,$5}' ejercicio_5.txt
DayofMonth DepTime
14 741
15 729
17 741
18 729
19 749
21 728
22 728
23 731
24 744
25 729
26 735
28 741
29 742
31 726
1 936
2 918
3 928
4 914
5 1042
6 934
7 946
8 932
9 947
10 915
13 1531
13 1910
13 1441
13 921
13 1435
13 1750
13 706
```

10.- Sin usar vim, nano o editor de texto alguno use comandos de Linux para reemplazar TODOS los elementos de la segunda columna por -1, guárdelo como **archivo_6.txt** y hágale un cat a ese mismo archivo.

Resultado: `awk -F ',' '{print $2}' ejercicio_5.txt > archivo_6.txt | sed -i 's/10/-1/g' archivo_6.txt`

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ awk -F ',' '{print $2}' ejercicio_5.txt > archivo_6.txt | sed -i 's/10/-1/g'
sed: no input files
[cloudera@quickstart datio_jctm]$ awk -F ',' '{print $2}' ejercicio_5.txt > archivo_6.txt | sed -i 's/10/-1/g' archivo_6.txt
[cloudera@quickstart datio_jctm]$ ls
aerolinea.csv aerolinea_head10.csv archivo_6.txt ejercicio_2.txt ejercicio_3.txt ejercicio_4.txt ejercicio_5.txt tee
[cloudera@quickstart datio_jctm]$
```

SECCION 2. HDFS Y HIVE

11.- Se está tratando de hacer la siguiente operación:

`hdfs dfs -head /raw/aerolínea.csv`


```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ hdfs dfs
Usage: hadoop fs [generic options]
    [-appendToFile <localsrc> ... <dst>]
    [-cat [-ignoreCrc] <src> ...]
    [-checksum <src> ...]
    [-chgrp [-R] GROUP PATH...]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-copyFromLocal [-f] [-p] [-l] <localsrc> ... <dst>]
    [-copyToLocal [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
    [-count [-q] [-h] [-v] [-x] <path> ...]
    [-cp [-f] [-p | -p[topax]] <src> ... <dst>]
    [-createSnapshot <snapshotDir> [<snapshotName>]]
    [-deleteSnapshot <snapshotDir> <snapshotName>]
    [-df [-h] [<path> ...]]
    [-du [-s] [-h] [-x] <path> ...]
    [-expunge]
    [-find <path> ... <expression> ...]
    [-get [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
    [-getfacl [-R] <path>]
    [-getfattr [-R] {-n name | -d} [-e en] <path>]
    [-getmerge [-nl] <src> <localdst>]
    [-help [cmd ...]]
    [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
    [-mkdir [-p] <path> ...]
    [-moveFromLocal <localsrc> ... <dst>]
    [-moveToLocal <src> <localdst>]
    [-mv <src> ... <dst>]
    [-put [-f] [-p] [-l] <localsrc> ... <dst>]
    [-renameSnapshot <snapshotDir> <oldName> <newName>]
    [-rm [-f] [-r|-R] [-skipTrash] <src> ...]
    [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
    [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
    [-setfattr {-n name [-v value] | -x name} <path>]
    [-setrep [-R] [-w] <rep> <path> ...]
    [-stat [format] <path> ...]
    [-tail [-f] <file>]
    [-test [-defsz] <path>]
    [-text [-ignoreCrc] <src> ...]
    [-touchz <path> ...]
    [-usage [cmd ...]]

Generic options supported are
-conf <configuration file>    specify an application configuration file
```

En el HDFS no se tiene la función o comando head por lo cual la ejecución de la línea “hdfs dfs -head /raw/aerolínea.csv” ejecuta un error

```
[cloudera@quickstart datio_jctm]$ hdfs dfs -head /raw/aerolinea.csv
-head: Unknown command
[cloudera@quickstart datio_jctm]$
```

Con una captura muestre qué es lo que pasa y por medio de argumentos sólidos (una captura de pantalla con la evidencia, una fuente de consulta) por qué sucede esto.

12.- Cuente cuántas líneas tiene el archivo **aerolínea.csv** que está **en el HDFS**. Recuerde el carácter pipe (|) empleado en ejercicios anteriores.

```
hdfs dfs -cat /raw/ aerolínea.csv | wc -l
```

13.- Indague en la instrucción de HDFS para averiguar el factor de réplica del archivo aerolínea.csv y colóquelo aquí junto con captura del resultado.

```
hdfs dfs -stat %r /jctmmx/aerolínea.csv
```



```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ hdfs dfs -stat %r /jctmmx/aerolinea.csv
1
[cloudera@quickstart datio_jctm]$
```

14.- Tome como base el archivo **aerolínea.csv** del HDFS y almacene en el sistema local un archivo **ejercicio_14.txt** que contenga las primeras 15 líneas sin usar el comando -tail del HDFS. Muestre ese contenido también.

```
hdfs dfs -cat /jctmmx/aerolínea.txt | head -n 15 | tee ejercicio_14.txt
```

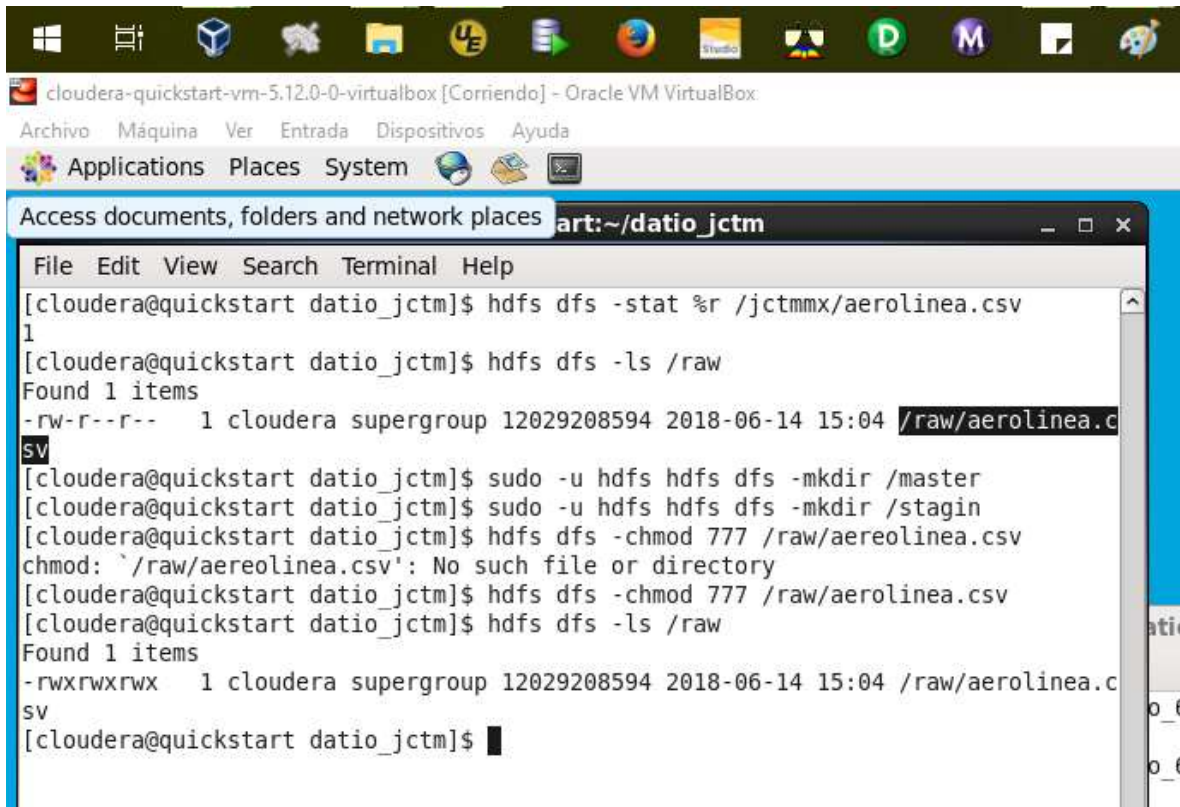
15.- Cree los directorios **master** y **stagin** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso. Coloque las capturas de ambos ejercicios por separado.

```
hdfs dfs -ls /raw
```

```
sudo -u hdfs hdfs dfs -mkdir /master
```

```
sudo -u hdfs hdfs dfs -mkdir /stagin
```

```
hdfs dfs -ls /raw
```

```
cloudera-quickstart-vm-5.12.0-0-virtualbox [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Applications Places System
Access documents, folders and network places art:~/datio_jctm
File Edit View Search Terminal Help
[cloudera@quickstart datio_jctm]$ hdfs dfs -stat %r /jctmmx/aerolinea.csv
1
[cloudera@quickstart datio_jctm]$ hdfs dfs -ls /raw
Found 1 items
-rw-r--r-- 1 cloudera supergroup 12029208594 2018-06-14 15:04 /raw/aerolinea.c
sv
[cloudera@quickstart datio_jctm]$ sudo -u hdfs hdfs dfs -mkdir /master
[cloudera@quickstart datio_jctm]$ sudo -u hdfs hdfs dfs -mkdir /stagin
[cloudera@quickstart datio_jctm]$ hdfs dfs -chmod 777 /raw/aereolinea.csv
chmod: `/raw/aereolinea.csv': No such file or directory
[cloudera@quickstart datio_jctm]$ hdfs dfs -chmod 777 /raw/aerolinea.csv
[cloudera@quickstart datio_jctm]$ hdfs dfs -ls /raw
Found 1 items
-rwxrwxrwx 1 cloudera supergroup 12029208594 2018-06-14 15:04 /raw/aerolinea.c
sv
[cloudera@quickstart datio_jctm]$
```

16.- Para los siguientes ejercicios puede hacer uso del servicio Hue (si no ha activado los servicios en Cloudera Manager tiene que hacerlo antes, para entrar a Hue en el mismo navegador se encuentra esta opción).

Aparecerá una ventana como ésta:

```
CREATE EXTERNAL TABLE tabla_aerolinea(
```

```
Year STRING, Month STRING, DayofMonth STRING, DayOfWeek STRING, DepTime STRING,
CRSDepTime STRING, ArrTime STRING, CRSArrTime STRING, UniqueCarrier STRING, FlightNum
STRING, TailNum STRING, ActualElapsedTime STRING, CRSElapsedTime STRING, AirTime STRING,
ArrDelay STRING, DepDelay STRING, Origin STRING, Dest STRING, Distance STRING,
```

```
TaxiIn STRING, TaxiOut STRING, Cancelled STRING, CancellationCode STRING, Diverted STRING,
```

```
CarrierDelay STRING, WeatherDelay STRING,
```

```
NASDelay STRING, SecurityDelay STRING, LateAircraftDelay STRING),
```

```
Adicional STRING
```

```
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE location '/raw'
```

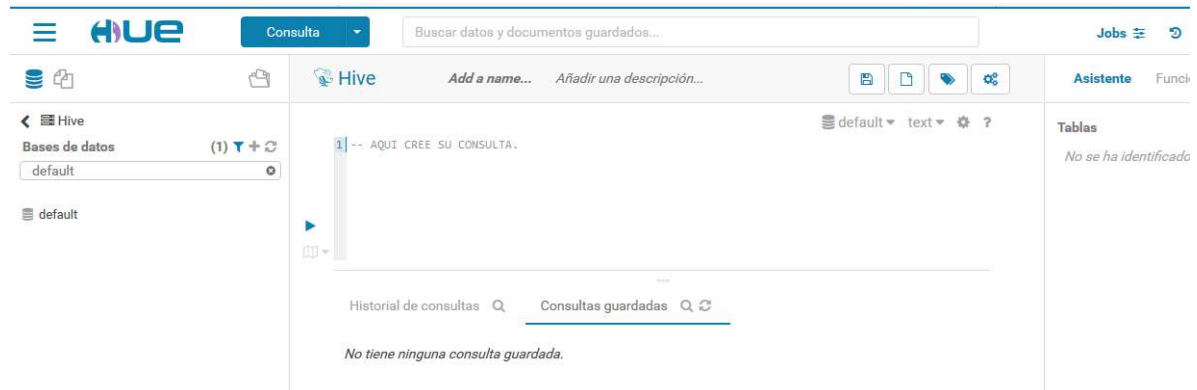
TBLPROPERTIES ('SKIP.HEADER.LINE.COUNT'='1');

The screenshot shows the Cloudera Hue web interface. At the top, there's a navigation bar with links to Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. Below this is a search bar and a 'Query' button. The main area is divided into three sections: a left sidebar with a table of contents, a central query results area, and a right sidebar with 'Assistant' and 'Functions' tabs. The central area displays a table with 10 rows of data, including columns for year, month, dayofmonth, dayofweek, deptime, crsdeptime, and arrtime. The right sidebar shows a list of tables, including 'default.tabla_aerolinea'.

| | year | month | dayofmonth | dayofweek | deptime | crsdeptime | arrtime |
|----|------|-------|------------|-----------|---------|------------|---------|
| 1 | Year | Month | DayOfMonth | DayOfWeek | DepTime | CRSDepTime | ArrTime |
| 2 | 1987 | 10 | 14 | 3 | 741 | 730 | 912 |
| 3 | 1987 | 10 | 15 | 4 | 729 | 730 | 903 |
| 4 | 1987 | 10 | 17 | 6 | 741 | 730 | 918 |
| 5 | 1987 | 10 | 18 | 7 | 729 | 730 | 847 |
| 6 | 1987 | 10 | 19 | 1 | 749 | 730 | 922 |
| 7 | 1987 | 10 | 21 | 3 | 728 | 730 | 848 |
| 8 | 1987 | 10 | 22 | 4 | 728 | 730 | 852 |
| 9 | 1987 | 10 | 23 | 5 | 731 | 730 | 902 |
| 10 | 1987 | 10 | 24 | 6 | 744 | 730 | 908 |

The image shows the Hue login page. It features the Hue logo at the top, followed by the tagline 'Query. Explore. Repeat.'. Below this are two input fields: 'Nombre de usuario' (Username) and 'Contraseña' (Password). A blue button labeled 'Iniciar sesión' (Log in) is positioned below the password field. At the bottom of the page, there is a footer that reads 'Hue y el logotipo de Hue son marcas comerciales de Cloudera, Inc.'

Recuerde que tanto el usuario como la contraseña es **cloudera**:



Entonces tome el siguiente código y cree una tabla en Hive:

```
CREATE EXTERNAL TABLE tabla_aerolinea(
```

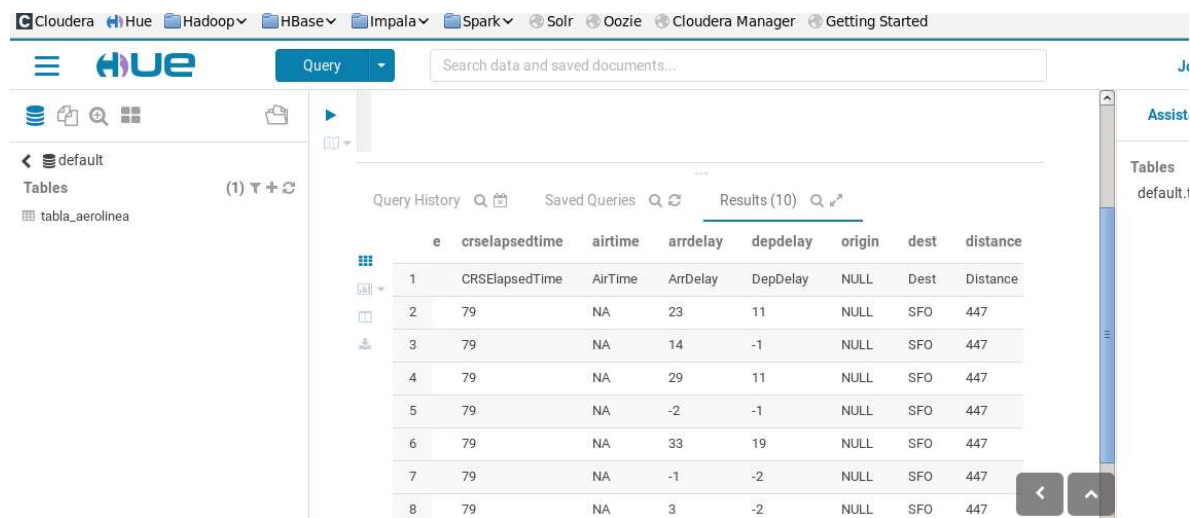
```
Year STRING,  
Month STRING,  
DayofMonth STRING,  
DayOfWeek STRING,  
DepTime STRING,  
CRSDepTime STRING,  
ArrTime STRING,  
CRSArrTime STRING,  
UniqueCarrier STRING,  
FlightNum STRING,  
TailNum STRING,  
ActualElapsedTime STRING,  
CRSElapsedTime STRING,  
AirTime STRING,  
ArrDelay STRING,  
DepDelay STRING,  
Origin STRING,  
Dest STRING,  
Distance STRING,  
TaxiIn STRING,  
TaxiOut STRING,
```

Cancelled STRING,
CancellationCode STRING,
Diverted STRING,
CarrierDelay STRING,
WeatherDelay STRING,
NASDelay STRING,
SecurityDelay STRING,
LateAircraftDelay STRING)

ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
location '/raw';

En el código anterior **NO** existe una forma de omitir los encabezados por lo que es su deber encontrar esa manera, incluirla en el código y crear la tabla.
Para acreditar el ejercicio debe mostrar la sentencia que requirió para la parte de los encabezados y hacer un SELECT de los 10 primeros elementos de la tabla.

17.- Borre la tabla anterior y vuélvala a crear pero ahora el tipo de dato Origin debe ser INT, entonces vuelva a ejecutar la consulta y especifique qué ha pasado y con una captura muéstrela.



| e | crselapsedtime | airtime | arrdelay | depdelay | origin | dest | distance |
|---|----------------|---------|----------|----------|--------|------|----------|
| 1 | CRSElapsedTime | AirTime | ArrDelay | DepDelay | NULL | Dest | Distance |
| 2 | 79 | NA | 23 | 11 | NULL | SFO | 447 |
| 3 | 79 | NA | 14 | -1 | NULL | SFO | 447 |
| 4 | 79 | NA | 29 | 11 | NULL | SFO | 447 |
| 5 | 79 | NA | -2 | -1 | NULL | SFO | 447 |
| 6 | 79 | NA | 33 | 19 | NULL | SFO | 447 |
| 7 | 79 | NA | -1 | -2 | NULL | SFO | 447 |
| 8 | 79 | NA | 3 | -2 | NULL | SFO | 447 |

18.- Borre la tabla anterior, vuélvala a crear (con Origin STRING) pero ahora añada una columna después de LateAircraftDelay llamada **Adicional** con tipo de dato **STRING**, ejecute la creación, indique qué ha sucedido y coloque captura del resultado.

The screenshot shows the Hue web interface. At the top, there's a search bar and a 'Query' dropdown. On the left, a sidebar shows the 'default' database and a table named 'tabla_aerolinea'. The main area displays a SQL query: `select * from tabla_aerolinea;`. Below the query, the results are shown in a table with 7 columns: 'lay', 'weatherdelay', 'nasdelay', 'securitydelay', 'lateaircraftdelay', and 'adicional'. The results table has 5 rows, all with 'NA' values except for the first row which has specific delay names. The interface also shows 'Query History', 'Saved Queries', and 'Results (1,024+)'. On the right, there's a 'Table de' sidebar.

| lay | weatherdelay | nasdelay | securitydelay | lateaircraftdelay | adicional |
|-----|--------------|----------|---------------|-------------------|-----------|
| 1 | WeatherDelay | NASDelay | SecurityDelay | LateAircraftDelay | NULL |
| 2 | NA | NA | NA | NA | NULL |
| 3 | NA | NA | NA | NA | NULL |
| 4 | NA | NA | NA | NA | NULL |
| 5 | NA | NA | NA | NA | NULL |

19.- En esta tabla anterior inserte un renglón a la tabla con todos los valores iguales a “NA” (tiene que investigar cómo añadir elementos a la tabla), y luego después de la inserción del elemento indague en qué parte del HDFS se ha guardado ese nuevo elemento.

SECCIÓN 3. PREGUNTAS ABIERTAS

20.- ¿Qué es el Sticky Bit? Ejemplifíquelo con el archivo **ejercicio_5.txt** adjuntando una captura de pantalla.

Permiso de acceso a los fichos

```
-rw-r--r-- 1 root    root      4.4K Jun 25 11:22 zookeeper.out
[cloudera@quickstart datio_jctm]$ chmod 1755 ejercicio_5.txt
[cloudera@quickstart datio_jctm]$ ls -lh
total 12G
-rwxrwxrwx 1 cloudera cloudera 12G Jun 12 16:22 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 14 15:18 aerolinea_head10.csv
-rw-rw-r-- 1 cloudera cloudera   0 Jun 25 10:20 archivo_6.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:49 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 07:53 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 25 09:20 ejercicio_4.txt
-rwxr-xr-t 1 cloudera cloudera 5.1K Aug 25 2018 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.1K Jun 25 09:27 tee
-rw-r--r-- 1 root    root      4.4K Jun 25 11:22 zookeeper.out
[cloudera@quickstart datio_jctm]$
```

21.- ¿A qué se le conoce como NoSQL?, ¿considera que Hive e Impala son representantes? Justifique la respuesta.

SQL para objetos que no son estructuradas de las BDR, no requieren tablas estrutrudadas, hive administra e impala ejecta procesos nosql

22.- Investigue el uso del comando nohup en GNU/Linux y con base en esto responda: ¿cómo puede ser aplicado dicho comando en un sistema distribuido?

Comando que permite la ejecución de los procesos sin perder de vista el proceso



The screenshot shows a terminal window titled "cloudera@quickstart:~/datio_jctm". The terminal displays the man page for the 'nohup' command. The page includes sections for NAME, SYNOPSIS, DESCRIPTION, and AUTHOR. The DESCRIPTION section explains that 'nohup' runs a command immune to hangups, with output to a non-tty. It also lists options like --help and --version. The AUTHOR section states it was written by Jim Meyering.

```
cloudera@quickstart:~/datio_jctm
File Edit View Search Terminal Help
NOHUP(1)                                User Commands                                NOHUP(1)

NAME
    nohup - run a command immune to hangups, with output to a non-tty

SYNOPSIS
    nohup COMMAND [ARG]...
    nohup OPTION

DESCRIPTION
    Run COMMAND, ignoring hangup signals.

    --help display this help and exit

    --version
        output version information and exit

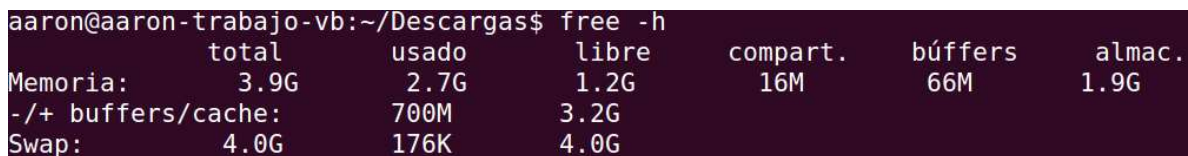
    If standard input is a terminal, redirect it from /dev/null. If stan-
    dard output is a terminal, append output to 'nohup.out' if possible,
    '$HOME/nohup.out' otherwise. If standard error is a terminal, redirect
    it to standard output. To save output to FILE, use 'nohup COMMAND >
    FILE'.

    NOTE: your shell may have its own version of nohup, which usually
    supersedes the version described here. Please refer to your shell's
    documentation for details about the options it supports.

AUTHOR
    Written by Jim Meyering.

REPORTING BUGS
    :
```

23.- Se quiere averiguar la memoria RAM disponible con base en la siguiente imagen:



The screenshot shows a terminal window with the command 'free -h' executed. The output displays memory usage statistics in human-readable format, including total, used, free, shared, buffers, and available memory.

```
aaron@aaron-trabajo-vb:~/Descargas$ free -h
              total        usado       libre       compart.     búffers       almac.
Memoria:      3.9G         2.7G         1.2G          16M          66M         1.9G
-/+ buffers/cache:  700M         3.2G
Swap:         4.0G         176K         4.0G
```

Indique el o los valores adecuados y por qué.

Es 3.9 G de los cuales usado tiene 2.7, el Swap es la parte de disco duro destinada para ser usada como RAM

24.- Se tiene el siguiente escenario: personal ajeno a su área de sistemas desea tener acceso al sistema, en particular para ver algunos datos del archivo **objetivo.txt**

Por otra parte se sabe de manera extraoficial que la meta de ellos consiste en “ensuciar” el archivo para que el área no tenga tanto repunte como la nuestra.

Por cuestiones burocráticas la creación de algún usuario nuevo no es plausible no obstante debido a asuntos políticos es prácticamente un hecho que se le tiene que dar permiso, por ello es que se optó por prestarles un usuario (**usuario_nuestro**) cuyo grupo es **grupo_nuestro**.

Con base en estas características y limitando el escenario únicamente a comandos **chmod (y si lo desea chown y chgrp)**, ¿cuál sería la configuración que usted propondría para garantizar el acceso al archivo pero al mismo tiempo protegerlo de las circunstancias mencionadas y sin afectar al mismo tiempo a los demás miembros de **grupo_nuestro**?

`chmod 755 objetivo.txt`

25.- ¿Cuál es la diferencia entre Hadoop y Cloudera?

Cloudera es la herramienta bigdata que incluye hadoop como manejador de archivos distribuidos y

26.- ¿Cuáles son los tipos de archivos existentes en GNU/Linux y Windows?

Unix/Linux

Ext, Ext2, Ext3, Swap

Windows

Fat32, FAT 16, NTFS

27.- ¿Qué es el SerDe y cuál es su relación con Hive e Impala?

Es una función de serializable y deserializable de los archivos y se utiliza en ambos, hive e impala

28.- ¿A qué se le conoce como Big Table y Big Query?

BigTable, es el sistema de Google de distribuido

BigQuery es el servicio web de Google para consultar

29.- ¿A qué se le denomina Data Lake y Data Warehouse?

Data Lake, es la herramienta para realizar la ingesta de archivos

Data Warehouse, sistema estructurado que permite el proceso de info

30.- ¿Existe algún otro tipo de sistemas de archivos distribuidos que NO sea HDFS? si es así, ¿de cuáles se trata?

Open GFS,

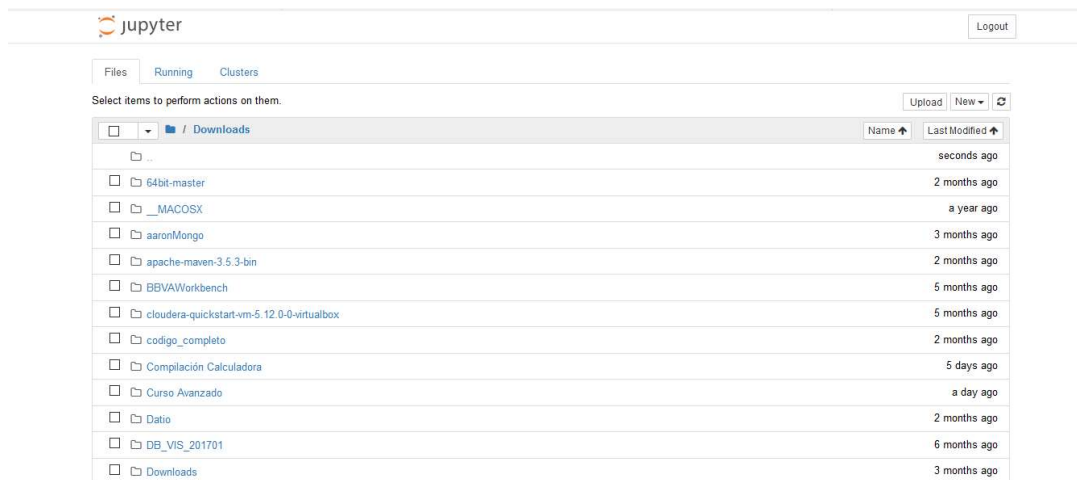
MooseFS.

SECCIÓN 4. ESPECIAL

31.- Instale Jupyter en Cloudera, para ello puede basarse en la siguiente liga:

<https://medium.com/@vando/install-jupyter-notebook-on-centos-7-1d596abf08da>

Es importante señalar que para continuar el curso es imprescindible esta herramienta y no existirán pausas para su instalación durante las sesiones, motivo por el cual es menester llevar a cabo esta operación aunque solamente valga 1 crédito. Para validar este ejercicio se requiere una captura de pantalla del menú principal, algo así:



sudo -u yum install -y http://dl.fedoraproject.org/pub/epel/7/x86_64/Packages/e/epel-release-7-11.noarch.rpm

sudo yum install -y python-pip python-devel python-virtualenv

sudo yum groupinstall 'Development Tools'

virtualenv jupyter-virtualenv

source jupyter-virtualenv/bin/activate

```
pip install jupyter
```