

## Tarea de Vacaciones

Para esta tarea se deberá responder una serie de preguntas de temas que se han abordado en el curso, el objetivo consiste entonces en reforzar los conocimientos e indagar en otros nuevos que, aunque no forman parte explícita del temario, le servirán al estudiante para incursionar en temas Big Data a plenitud.

El estudiante debe crear primero que nada un directorio dentro de su directorio local de Git llamado **TareaVacaciones** y dentro de éste crear una copia de esta tarea que lleve por nombre **TareaX y colocarla juntos con los resultados en formato PDF**, donde X es su nombre de usuario empleado en Github. Por ejemplo:

### TareaYoNoFui

Con respecto de los ejercicios, a menos que se indique lo contrario, todas las respuestas constarán del código o instrucción resultante acompañada de una captura de pantalla. Ejemplo:

-1.- Indique el comando que se emplea para listar archivos en GNU/Linux de manera simple:

Respuesta: **ls**

```
aaron@aaron-trabajo-vb:~/Descargas$ ls
apache-hive-2.3.3-bin.tar.gz      R-3.3.2.tar.gz
archivo.txt                      rattle_5.0.18.tar.gz
banner-principal2.png           Respaldo Usuaría Chile
core-site.xml                   RGtk2_2.20.33.tar.gz
db-derby-10.13.1.1-bin.tar.gz   rstudio-1.0.143-amd64.deb
file01.txt                      sas2txt.py
file02.txt                      scala-2.10.4.deb
file03.txt                      scala-2.12.1.tgz
```

En este tipo de preguntas de faltar alguno de los elementos señalados se considerará como errónea la respuesta y no se obtendrá el acierto.

Es menester mencionar que hay casos donde las preguntas son de tipo abierto, entonces en esos casos lo único que se pide adjuntar es tanto la respuesta como la(s) fuente(s). Ejemplo:

0.- ¿Cuál es el significado de la vida?

Respuesta: **42**

Fuente: <https://www.independent.co.uk/life-style/history/42-the-answer-to-life-the-universe-and-everything-2205734.html>

De nueva cuenta, si no existe al menos uno de estos dos elementos, la respuesta se considerará como inválida.

La fecha límite de entrega es el **Lunes 25 de Junio a las 15:00:00**, como se había mencionado con anterioridad el flujo de archivos se mantiene única y exclusivamente por Github, para ello se dejan los comandos a emplearse:

- **git pull** (para actualizar el repositorio)
- **git add .** (para indicar todos los elementos que se desean agregar al repositorio)
- **git commit -m "TareaVacaciones nombre\_usuario"** (para colocar un mensaje que distingua a esta subida de las de los demás usuarios)
- **git push origin master** (para efectuar los cambios)

Por cierto que en lo que se repara Git en Cloudera puede ocupar Git de Windows y de esta manera, ya que se sugirió la instalación de Guest Additions en VirtualBox, copiar los resultados al primer sistema.

Nuevamente, de no cumplirse al menos uno de los señalamientos anteriores la tarea se considerará no entregada.

Dicho lo anterior se les desea mucho éxito en la travesía, cualquier cosa no duden en preguntar...

## 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 25q aerolinea.csv

```
[cloudera@quickstart ~]$ sed 25q aerolinea.csv
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,58,46,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 ~]$
```

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.lininfo.org/pipes.html>

**Respuesta:** sed 25q aerolinea.csv | tee ejercicio\_2.txt

```
1987,10,10,0,913,913,1022,1001,PS,1451,NA,0,40,NA,21,0,SFO,RNO,192,NA,NA,0,NA,NA,NA,NA,NA
[cloudera@quickstart ~]$ sed 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,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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
[cloudera@quickstart ~]$
```

3.- Cambie el nombre del archivo **ejercicio\_2.txt** a **ejercicio\_3.txt** **SIN** usar el comando rename

**Respuesta:** mv ejercicio\_2.txt ejercicio3\_txt

```
1987,10,10,0,913,913,1022,1001,PS,1451,NA,0,40,NA,21,0,SFO,RNO,192,NA,NA,0,NA,NA,NA,NA,NA
[cloudera@quickstart ~]$ mv ejercicio_2.txt ejercicio_3.txt
[cloudera@quickstart ~]$
```

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**

**Respuesta:** tac aerolinea.csv | head -10 | tac > ejercicio\_4.txt

```
2000,12,13,0,1007,047,1149,1010,DL,1031,N909DA,102,143,122,99,00,ATL,IAN,
[cloudera@quickstart ~]$ tac aerolinea.csv | head -10 > ejercicio_4.txt
[cloudera@quickstart ~]$
```

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**

**Respuesta:** cat ejercicio\_3.txt ejercicio\_4.txt | tee ejercicio\_5.txt

```
[cloudera@quickstart ~]$ cat ejercicio_3.txt ejercicio_4.txt | tee ejercicio_5.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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,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,NA
2008,12,13,6,1110,1103,1413,1418,DL,1641,N908DL,123,135,104,-5,7,SAT,ATL,874,8,11,0,,0,NA,NA,NA,NA,NA,NA
2008,12,13,6,1251,1240,1446,1437,DL,1639,N646DL,115,117,89,9,11,IAD,ATL,533,13,13,0,,0,NA,NA,NA,NA,NA,NA
2008,12,13,6,655,700,856,856,DL,1638,N671DN,121,116,85,0,-5,PBI,ATL,545,24,12,0,,0,NA,NA,NA,NA,NA,NA
```

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.

**Respuesta:** **ls -l -h**

```
[cloudera@quickstart ~]$ ls -l -h
total 12G
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 13 16:41 aerolinea2.csv
-rwxrwxrwx 1 cloudera cloudera 12G Jun 13 11:40 aerolinea.csv
-rwxrwxr-x 1 cloudera cloudera 5.3K Jul 19 2017 cloudera-manager
-rwxrwxr-x 1 cloudera cloudera 9.8K Jul 19 2017 cm_api.py
drwxrwxr-x 2 cloudera cloudera 4.0K Jun 13 16:14 Desktop
drwxrwxr-x 4 cloudera cloudera 4.0K Jul 19 2017 Documents
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Downloads
drwxrwsr-x 9 cloudera cloudera 4.0K Feb 19 2015 eclipse
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 20 09:32 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 984 Jun 20 10:11 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 3.5K Jun 20 10:12 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 53K Jul 19 2017 enterprise-deployment.json
-rw-rw-r-- 1 cloudera cloudera 50K Jul 19 2017 express-deployment.json
-rwxrwxr-x 1 cloudera cloudera 4.9K Jul 19 2017 kerberos
drwxrwxr-x 2 cloudera cloudera 4.0K Jul 19 2017 lib
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Music
-rwxrwxr-x 1 cloudera cloudera 4.2K Jul 19 2017 parcels
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Pictures
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Public
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Templates
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Videos
drwxrwxr-x 5 cloudera cloudera 4.0K Jun 13 16:01 workspace
[cloudera@quickstart ~]$
```

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.

**Respuesta:** **touch -d '25 Aug' ejercicio\_5.txt**

```

[cloudera@quickstart ~]$ touch -d '25 Aug' ejercicio_5.txt
[cloudera@quickstart ~]$ ls -l -h
total 12G
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 13 16:41 aerolinea2.csv
-rwxrwxrwx 1 cloudera cloudera 12G Jun 13 11:40 aerolinea.csv
-rwxrwxr-x 1 cloudera cloudera 5.3K Jul 19 2017 cloudera-manager
-rwxrwxr-x 1 cloudera cloudera 9.8K Jul 19 2017 cm_api.py
drwxrwxr-x 2 cloudera cloudera 4.0K Jun 13 16:14 Desktop
drwxrwxr-x 4 cloudera cloudera 4.0K Jul 19 2017 Documents
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Downloads
drwxrwsr-x 9 cloudera cloudera 4.0K Feb 19 2015 eclipse
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 20 09:32 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 984 Jun 20 10:11 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 3.5K Aug 25 2018 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 53K Jul 19 2017 enterprise-deployment.js
-rw-rw-r-- 1 cloudera cloudera 50K Jul 19 2017 express-deployment.json
-rwxrwxr-x 1 cloudera cloudera 4.9K Jul 19 2017 kerberos
drwxrwxr-x 2 cloudera cloudera 4.0K Jul 19 2017 lib
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Music
-rwxrwxr-x 1 cloudera cloudera 4.2K Jul 19 2017 parcels
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Pictures
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Public
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Templates
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Videos
drwxrwxr-x 5 cloudera cloudera 4.0K Jun 13 16:01 workspace
[cloudera@quickstart ~]$ █

```

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.

**Respuesta:** nproc --all

**Fuente :** <http://www.daniloaz.com/es/como-saber-cuantos-procesadores-y-nucleos-tiene-una-maquina-linux/>

```

[cloudera@quickstart ~]$ nproc --all
1

```

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:

```
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$
```

**Respuesta:** `awk 'BEGIN {FS = "," } {print $3,"",$5}' ejercicio_5.txt`

**Fuente:** <http://francisconi.org/linux/comandos/awk>

---

```
[cloudera@quickstart ~]$ awk 'BEGIN {FS = "," } {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 , 1110
13 , 1251
13 , 655
13 , 834
13 , 1002
13 , 749
13 , 612
```

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.

**Respuesta:** `sed 's/10/-1/' ejercicio_5.txt | sed 's/12/-1/' > ejercicio_6.txt`

```

[cloudera@quickstart ~]$ sed 's/10/-1/' ejercicio.5.txt | sed 's/12/-1/' > ejercicio.6.txt
[cloudera@quickstart ~]$ cat ejercicio.6.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,-1,14,3,741,730,9-1,849,PS,1451,NA,91,79,NA,23,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,15,4,729,730,903,849,PS,1451,NA,94,79,NA,14,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,17,6,741,730,918,849,PS,1451,NA,97,79,NA,29,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,18,7,729,730,847,849,PS,1451,NA,78,79,NA,-2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,19,1,749,730,922,849,PS,1451,NA,93,79,NA,33,19,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,21,3,728,730,848,849,PS,1451,NA,80,79,NA,-1,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,24,6,744,730,908,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,28,3,741,725,919,855,PS,1451,NA,98,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,5,1,1042,915,1-19,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,6,2,934,915,1024,1001,PS,1451,NA,58,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,7,3,946,915,1037,1001,PS,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,8,4,932,915,1033,1001,PS,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,9,5,947,915,1036,1001,PS,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
1987,-1,10,6,915,915,1022,1001,PS,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA
2008,-1,13,6,11-1,1103,1413,1418,DL,1641,N908DL,123,135,104,-5,7,SAT,ATL,874,8,11,0,0,NA,NA,NA,NA
2008,-1,13,6,1251,1240,1446,1437,DL,1639,N646DL,115,117,89,9,11,IAD,ATL,533,13,13,0,0,NA,NA,NA,NA
2008,-1,13,6,655,700,856,856,DL,1638,N671DN,121,116,85,0,-5,PBI,ATL,545,24,12,0,0,NA,NA,NA,NA

```

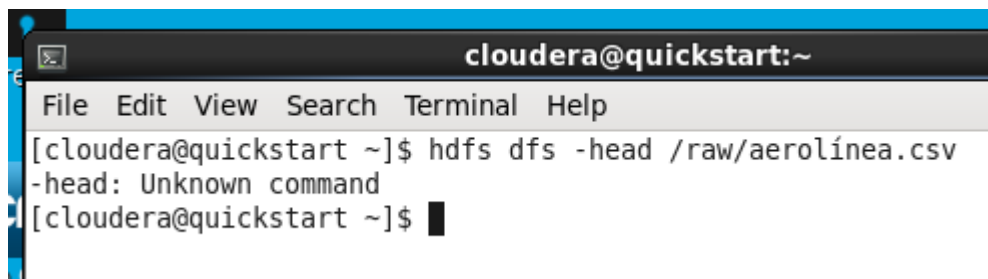
## SECCION 2. HDFS Y HIVE

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

**hdfs dfs -head /raw/aerolínea.csv**

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.

**Respuesta:** No existe el comando “head”



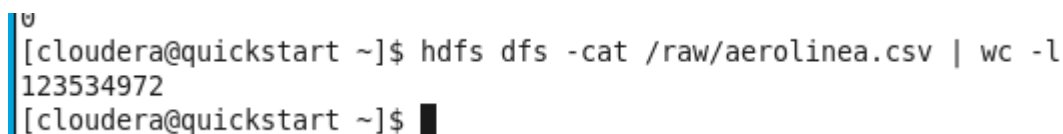
```

[cloudera@quickstart ~]$ hdfs dfs -head /raw/aerolínea.csv
-head: Unknown command
[cloudera@quickstart ~]$

```

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

**Respuesta:** hdfs dfs -cat /raw/aerolinea.csv | wc -l



```

[cloudera@quickstart ~]$ hdfs dfs -cat /raw/aerolinea.csv | wc -l
123534972
[cloudera@quickstart ~]$

```

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.

**Respuesta:** hdfs dfs -ls /raw/



**Fuente:** <https://www.systutorials.com/241375/how-to-check-the-replication-factor-of-a-file-in-hdfs/>

```
^C^C[cloudera@quickstart ~]$ hdfs dfs -ls /raw/
Found 3 items
drwxr-xr-x - cloudera supergroup 0 2018-06-22 16:36 /raw/.hive-staging_hive_2018-06-22_16-29-34_962_9118162406743353422-2
-rwxr-xr-x 1 cloudera supergroup 90 2018-06-22 16:36 /raw/000000_0
-rwxrw---- 1 cloudera supergroup 12029208594 2018-06-13 16:49 /raw/aerolinea.csv
```

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.

**Respuesta:** `hdfs dfs -cat /raw/aerolinea.csv | sed 15q | tee ejercicio_14.txt`

```
cat: Unable to write to output stream.
[cloudera@quickstart ~]$ hdfs dfs -cat /raw/aerolinea.csv | sed 15q | tee ejercicio_14.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
cat: Unable to write to output stream.
[cloudera@quickstart ~]$
```

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.

**Respuesta :** `sudo -u hdfs hdfs dfs -mkdir /master`

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

```
[cloudera@quickstart ~]$ sudo -u hdfs hdfs dfs -mkdir /master
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2017-07-19 05:34 /benchmarks
drwxr-xr-x - hbase supergroup 0 2018-06-22 13:57 /hbase
drwxr-xr-x - hdfs supergroup 0 2018-06-22 14:46 /master
drwxr-xr-x - cloudera supergroup 0 2018-06-13 16:49 /raw
drwxr-xr-x - solr solr 0 2017-07-19 05:37 /solr
drwxrwxrwt - hdfs supergroup 0 2018-06-13 16:24 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /user
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /var
```

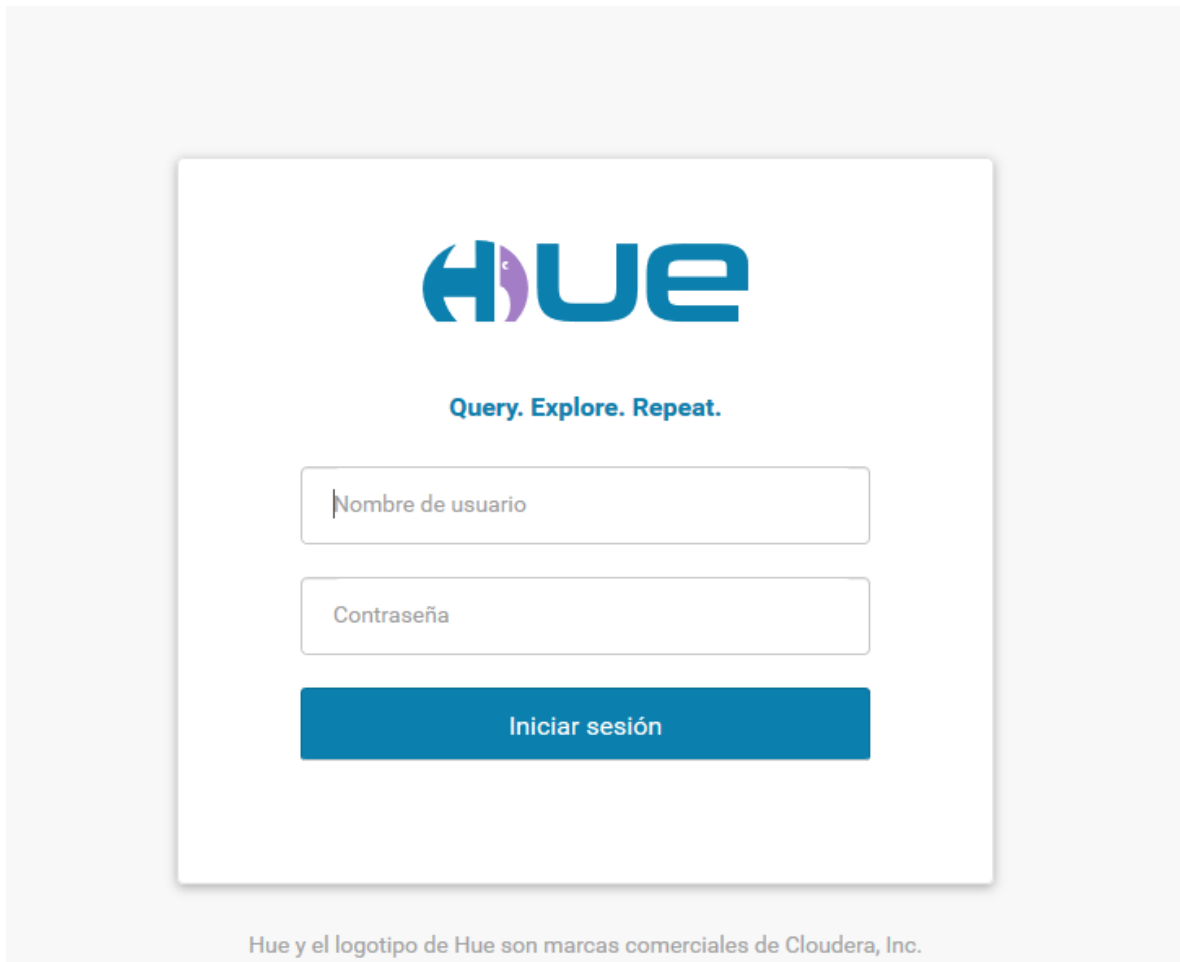
`sudo -u hdfs hdfs dfs -chmod 760 /raw/aerolinea.csv`

```
[cloudera@quickstart ~]$ hdfs dfs -ls /raw
Found 1 items
-rwxrw---- 1 cloudera supergroup 12029208594 2018-06-13 16:49 /raw/aerolinea.csv
[cloudera@quickstart ~]$
```

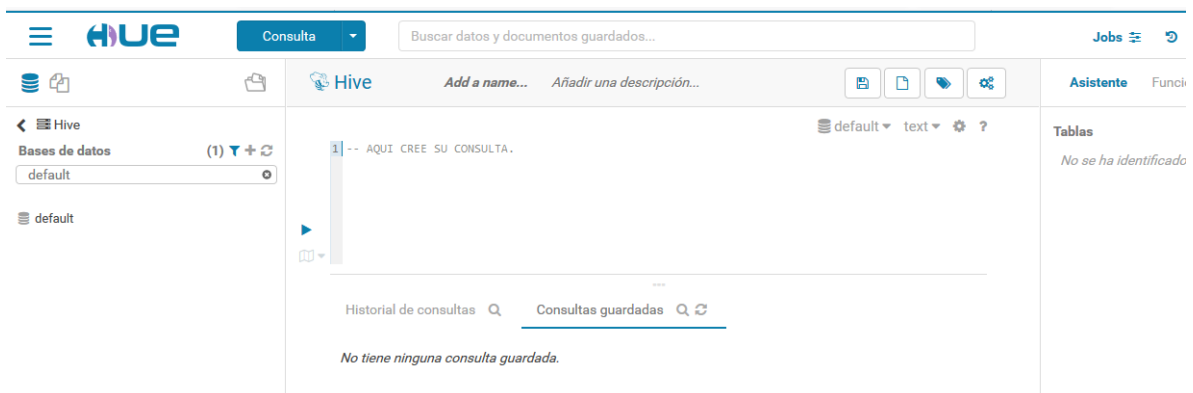


16.- Para los siguientes ejercicios puede hacer uso del servicio Hue (si no ha activado los servicios en Cloudera Manager tiene que hacerlo antes, para entrar a Hue en el mismo navegador se encuentra esta opción).

Aparecerá una ventana como ésta:



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.

**Respuesta:** tblproperties ("skip.header.line.count"="1")

```

27 WeatherDelay STRING,
28 NASDelay STRING,
29 SecurityDelay STRING,
30 LateAircraftDelay STRING)
31 ROW FORMAT DELIMITED
32 FIELDS TERMINATED BY ','
33 STORED AS TEXTFILE
34 location '/raw'
35 tblproperties ("skip.header.line.count"="1")
36 ;

```

Success.

1 | SELECT \* FROM tabla\_aerolinea LIMIT 10;

Query History | Saved Queries | Results (10)

	la_aerolinea.arrdelay	tabla_aerolinea.depdelay	tabla_aerolinea.origin	tabla_aer
1		11	SAN	SFO
2		-1	SAN	SFO
3		11	SAN	SFO
4		-1	SAN	SFO
5		19	SAN	SFO
6		-2	SAN	SFO

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 muéstrelo.

**Repuesta:** Se modificó el tipo de dato para Origin.

```

12 TailNum STRING,
13 ActualElapsedTime STRING,
14 CRSElapsedTime STRING,
15 AirTime STRING,
16 ArrDelay STRING,
17 DepDelay STRING,
18 Origin INT,

```

Success.

**Respuesta:** Se permite agregar el campo, pero no contiene valores.

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.

[illegible]

**Respuesta:** `hdfs dfs -ls /raw/`

### SECCIÓN 3. PREGUNTAS ABIERTAS

20.- ¿Qué es el Sticky Bit? Ejemplifíquelo con el archivo **ejercicio\_5.txt** adjuntando una captura de pantalla.

**Respuesta:** se utiliza para permitir que cualquiera pueda escribir y modificar sobre un archivo o directorio, pero que solo su propietario o root pueda eliminarlo

**Fuente:** <https://hvivani.com.ar/2013/09/06/permisos-especiales-sticky-bit-suid-sgid/>

```
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ chmod 1755 ejercicio_5.txt
chmod: cannot access `ejercicio_5.txt': No such file or directory
[cloudera@quickstart ~]$ chmod 1755 ejercicio_5.txt
[cloudera@quickstart ~]$ ls -lh
total 12G
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 13 16:41 aerolinea2.csv
-rwxrwxrwx 1 cloudera cloudera 12G Jun 13 11:40 aerolinea.csv
drwxrwxr-x 23 cloudera cloudera 4.0K Jun 21 08:49 anaconda3
-rwxrwxrwx 1 cloudera cloudera 622M Jun 20 15:15 Anaconda3-5.2.0-Lin
-rwxrwxr-x 1 cloudera cloudera 5.3K Jul 19 2017 cloudera-manager
-rwxrwxr-x 1 cloudera cloudera 9.8K Jul 19 2017 cm_api.py
drwxrwxr-x 2 cloudera cloudera 4.0K Jun 13 16:14 Desktop
drwxrwxr-x 4 cloudera cloudera 4.0K Jul 19 2017 Documents
drwxr-xr-x 2 cloudera cloudera 4.0K Jun 12 15:37 Downloads
drwxrwsr-x 9 cloudera cloudera 4.0K Feb 19 2015 eclipse
-rw-rw-r-- 1 cloudera cloudera 1.6K Jun 22 14:42 ejercicio_14.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 20 09:32 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 984 Jun 20 10:11 ejercicio_4.txt
-rwxr-xr-t 1 cloudera cloudera 3.5K Aug 25 2018 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3.5K Jun 20 15:26 ejercicio_6.txt
```

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

**Respuesta:** Las bases de datos NoSQL son sistemas de almacenamiento de información que no cumplen con el esquema entidad-relación. Tampoco utilizan una estructura de datos en forma de tabla donde se van almacenando los datos sino que para el almacenamiento hacen uso de otros formatos como clave-valor, mapeo de columnas o grafos.

**Apache Hive** es una infraestructura de almacenamiento de datos construida sobre Hadoop para proporcionar agrupación, consulta, y análisis de datos.<sup>1</sup> Inicialmente desarrollado por Facebook, Apache Hive es ahora utilizada y desarrollado por otras empresas como Netflix y la Financial Industry Regulatory Authority (FINRA).<sup>23</sup> Amazon mantiene una derivación de software de Apache Hive incluida en Amazon Elastic MapReduce en sus servicios Amazon Web Services.

Cloudera **Impala** es un motor de consulta que corre en Apache Hadoop. está dirigido a los analistas y científicos de datos para realizar análisis en los datos almacenados en Hadoop a través

de herramientas de SQL o business intelligence. El resultado es que el procesamiento de datos a gran escala (a través de MapReduce) y las consultas interactivas se pueden hacer en el mismo sistema utilizando los mismos datos y metadatos - eliminando la necesidad de migrar los conjuntos de datos a sistemas especializados y/o formatos propietarios solo para realizar el análisis.

**Fuente:**

- <https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf>
- [https://es.wikipedia.org/wiki/Apache\\_Hive](https://es.wikipedia.org/wiki/Apache_Hive)
- [https://es.wikipedia.org/wiki/Cloudera\\_Impala](https://es.wikipedia.org/wiki/Cloudera_Impala)

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?

**Respuesta :** permite mantener la ejecución de un comando pese a salir de la terminal, ya que hace que se ejecute de forma independiente a la sesión. Es útil en casos en los que como es habitual nos conectamos por conexión ssh a un host, si vamos a ejecutar un script que tarde bastante tiempo, o por el contrario no podemos correr el riesgo de que nos falle la red por diversos motivos y al finalizarse la conexión ssh, finalice los procesos que teníamos ligados a ella mandando la señal HUP de ahí el su nombre además de proporcionar un fichero de log de la ejecución

**Fuente:**

- <http://rm-rf.es/nohup-mantiene-ejecucion-comando-pese-salir-terminal/>
- <http://www.nexolinux.com/bash-scripting-nohup-y-procesos-en-segundo-plano/>

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

```
aaron@aaron-trabajo-vb:~/Descargas$ free -h
```

	total	usado	libre	compart.	buffers	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.

**Respuesta:** Se cuenta con 1.2G de espacio libre; al cual se le suman 2G de espacio que fue liberado y que en algún momento fue utilizado; teniendo un total de 3.2 G.

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**?

**Respuesta:** Asignaría permisos únicamente de lectura para los miembros usuarios del mismo grupo; todos los permisos para el propietario, tomando en cuenta que este usuario no será el que se preste al personal de otra área; y ningún permiso para cualquier otro:

Chmod 740 objetivo.txt

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

Cloudera es una firma especializada en Big Data, que permite añadir funciones a la arquitectura Hadoop de seguridad, control y gestión necesarios para establecer una solución empresarial robusta y fiable.

Su software está basado en Apache Hadoop y ofrecen soporte, servicios y formación para grandes clientes.

Hadoop es un ecosistema de componentes de código abierto que cambia fundamentalmente la manera en que las empresas almacenan, procesan y analizan datos.

A diferencia de los sistemas tradicionales, Hadoop permite que múltiples tipos de cargas de trabajo y analíticas se ejecuten en los mismos datos al mismo tiempo a gran escala.

**Referencia:** <https://www.clarcat.com/cloudera>

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

Linux:

- .bz2 — archivo comprimido con bzip2
- .gz — archivo comprimido con gzip
- .tar — archivo guardado con tar (iniciales de tape archive)
- .tbz — archivo tar y bzip
- .tgz — archivo tar y gzip.
- .zip — archivo comprimido con ZIP
- .au — archivo de audio
- .gif — archivo gráfico o de imagen
- .html/.htm — archivo HTML
- .jpg — archivo de imagen JPEG
- .pdf — imagen electrónica de un documento; PDF son las siglas de Portable Document Format



- .png — archivo gráfico de imagen PNG (siglas de Portable Network Graphic)
- .ps — archivo PostScript; formateado para ser impreso
- .txt — archivo plano de texto ASCII
- .wav — archivo de audio
- .xpm — archivo de imagen
- .conf — archivo de configuración. A veces los archivos de configuración usan la extensión .cfg, también.
- .lock — archivo lock; determina si el programa o dispositivo está en uso
- .rpm — archivo del gestor de paquetes de Red Hat que se usa para instalar software
- .c — archivo del código fuente del lenguaje de programación C
- .cpp — archivo del código fuente del lenguaje de programación C++
- .h — archivo cabecera de lenguaje de programación C o C++
- .o — archivo objeto de programación
- .pl — script Perl
- .py — script Python
- .so — archivo de librería
- .sh — script shell
- .tcl — script TCL

## Windows

- .docx Microsoft Open Word XML Document
- .doc Microsoft Word Document
- .txt Plain Text File
- .rtf Revit Family Template File
- .odt OpenOffice/StarOffice File
- .mp3 MP3 Audio File
- .wav Wave Audio File
- .aac MPEG-2 Advanced Audio Coding File
- .wma Windows Media Audio File
- .m4a MPEG-4 Audio File
- .avi Audio Video Interleave File
- .mp4 MPEG-4 Video File
- .mov Video Clip
- .flv Video File
- .mpg MPEG 1 System Stream
- .pdf Portable Document Format File
- .xls Excel Spreadsheet File
- .csv Comma Separated Values File
- .ini Initialization/Configuration File
- .html Hypertext Markup Language File

- .zip ZIP File
- .rar WinRAR Compressed Archive
- .7z Compressed File
- .tar Consolidated Unix File Archive
- .gz GNU Zipped Archive File
- .exe Windows Executable File
- .msi Windows Installer File
- .bin Binary Disc Image
- .app Punch Post
- .dmg Disk Copy Disk Image File

**Referencia:**

- <http://www.reviversoft.com/es/file-extensions/>
- <http://www.fis.unipr.it/pub/linux/redhat/9/en/doc/RH-DOCS/rhl-gsg-es-9/s1-managing-file-types.html>

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

SerDe es la abreviatura de Serializador / Deserializador; le permite a Hive leer los datos de una tabla y escribirlos nuevamente en HDFS en cualquier formato personalizado. Cualquiera puede escribir su propio SerDe para sus propios formatos de datos.

**Fuente:** <https://cwiki.apache.org/confluence/display/Hive/SerDe>

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

Big table BigTable es un sistema de gestión de base de datos creado por Google con las características de ser: distribuido, de alta eficiencia y propietario. Está construido sobre GFS(Google File System), Chubby Lock Service, y algunos otros servicios y programas de Google, y funciona sobre 'commodity hardware' (sencillos y baratos PCs con procesadores Intel).

BigTable almacena la información en tablas multidimensionales cuyas celdas están, en su mayoría, sin utilizar. Además, estas celdas disponen de versiones temporales de sus valores, con lo que se puede hacer un seguimiento de los valores que han tomado históricamente

BigQuery es un servicio web RESTful que permite el análisis interactivo de grandes conjuntos de datos que trabajan en conjunto con Google Storage. Es una Infraestructura como Servicio (IaaS) que puede usarse de forma complementaria con MapReduce

**Fuente:**

- <https://es.wikipedia.org/wiki/BigTable>
- <https://en.wikipedia.org/wiki/BigQuery>

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

Un data lake es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. A diferencia de un data warehouse jerárquico que almacena datos en ficheros o carpetas, un data lake utiliza una arquitectura plana para almacenar los datos.

Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

Fuente:

- <https://www.powerdata.es/data-warehouse#warehouse-1>
- <https://www.powerdata.es/data-lake#lake-1>

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

Quantcast File System (QFS) es un sistema de archivos distribuidos de alto rendimiento, de fuente abierta, tolerante a fallas, desarrollado para admitir el procesamiento de MapReduce u otras aplicaciones que leen y escriben grandes archivos secuencialmente

Ceph es un sistema de almacenamiento distribuido, masivo, de código abierto, unificado y escalable diseñado para un excelente rendimiento, confiabilidad y escalabilidad.

Lustre es un sistema de archivos masivamente distribuido, paralelo y global, generalmente utilizado para la computación en clúster a gran escala

GlusterFS es un sistema de archivos distribuido de código abierto capaz de escalar hasta varios petabytes y manejar miles de clientes.

Fuente:

- <https://www.linuxlinks.com/QuantcastFileSystem/>
- <https://www.linuxlinks.com/Lustre/>
- <https://www.linuxlinks.com/GlusterFS/>
- <https://www.linuxlinks.com/Ceph/>

#### 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:**

jupyter

Logout

FilesRunningClusters

Select items to perform actions on them.

UploadNew

0 / Downloads

	Name	Last Modified
<input type="checkbox"/>	...	seconds ago
<input type="checkbox"/>	64bit-master	2 months ago
<input type="checkbox"/>	_MACOSX	a year ago
<input type="checkbox"/>	aaronMongo	3 months ago
<input type="checkbox"/>	apache-maven-3.5.3-bin	2 months ago
<input type="checkbox"/>	BBVAWorkbench	5 months ago
<input type="checkbox"/>	cloudera-quickstart-vm-5.12.0-0-virtualbox	5 months ago
<input type="checkbox"/>	codigo_completo	2 months ago
<input type="checkbox"/>	Compilación Calculadora	5 days ago
<input type="checkbox"/>	Curso Avanzado	a day ago
<input type="checkbox"/>	Datio	2 months ago
<input type="checkbox"/>	DB_VIS_201701	6 months ago
<input type="checkbox"/>	Downloads	3 months ago

ArchivoMáquinaVerEntradaDispositivosAyuda

ApplicationsPlacesSystem

cloudera

Home - Mozilla Firefox

Home

localhost:8889/tree

Search

ClouderaHueHadoopHBaseImpalaSparkSolrOozieCloudera Manager

jupyter

QuitLogout

FilesRunningClusters

Select items to perform actions on them.

UploadNew

0 /

	Name	Last Modified	File size
<input type="checkbox"/>	anaconda3	3 minutes ago	
<input type="checkbox"/>	Desktop	8 days ago	
<input type="checkbox"/>	Documents	a year ago	
<input type="checkbox"/>	Downloads	9 days ago	
<input type="checkbox"/>	eclipse	3 years ago	
<input type="checkbox"/>	lib	a year ago	
<input type="checkbox"/>	Music	9 days ago	
<input type="checkbox"/>	Pictures	9 days ago	

cloudera 25 items, Free space: 14.6 GB