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***

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 distinga 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.

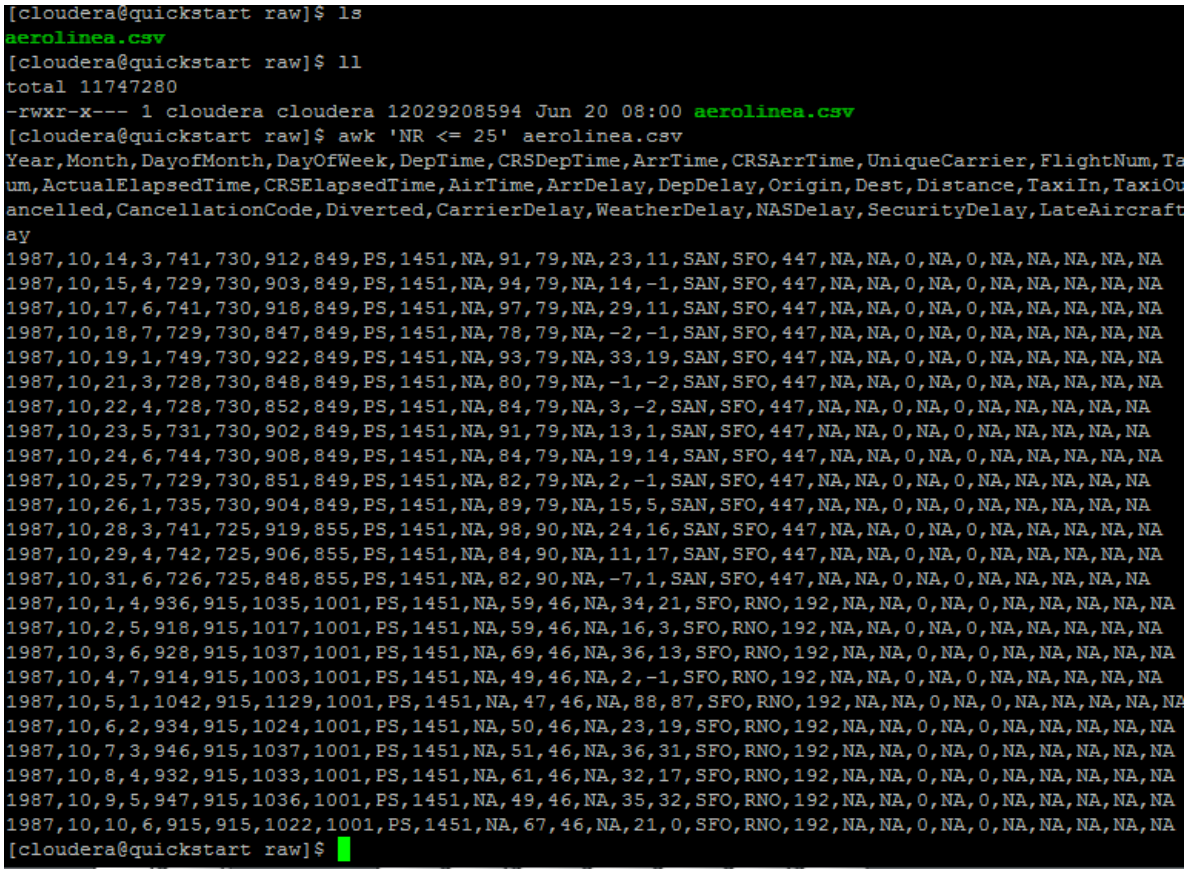
Nuevamente, de no cumplirse al menos uno de los señalamientos anteriores la tarea se considerará no entregada.

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.

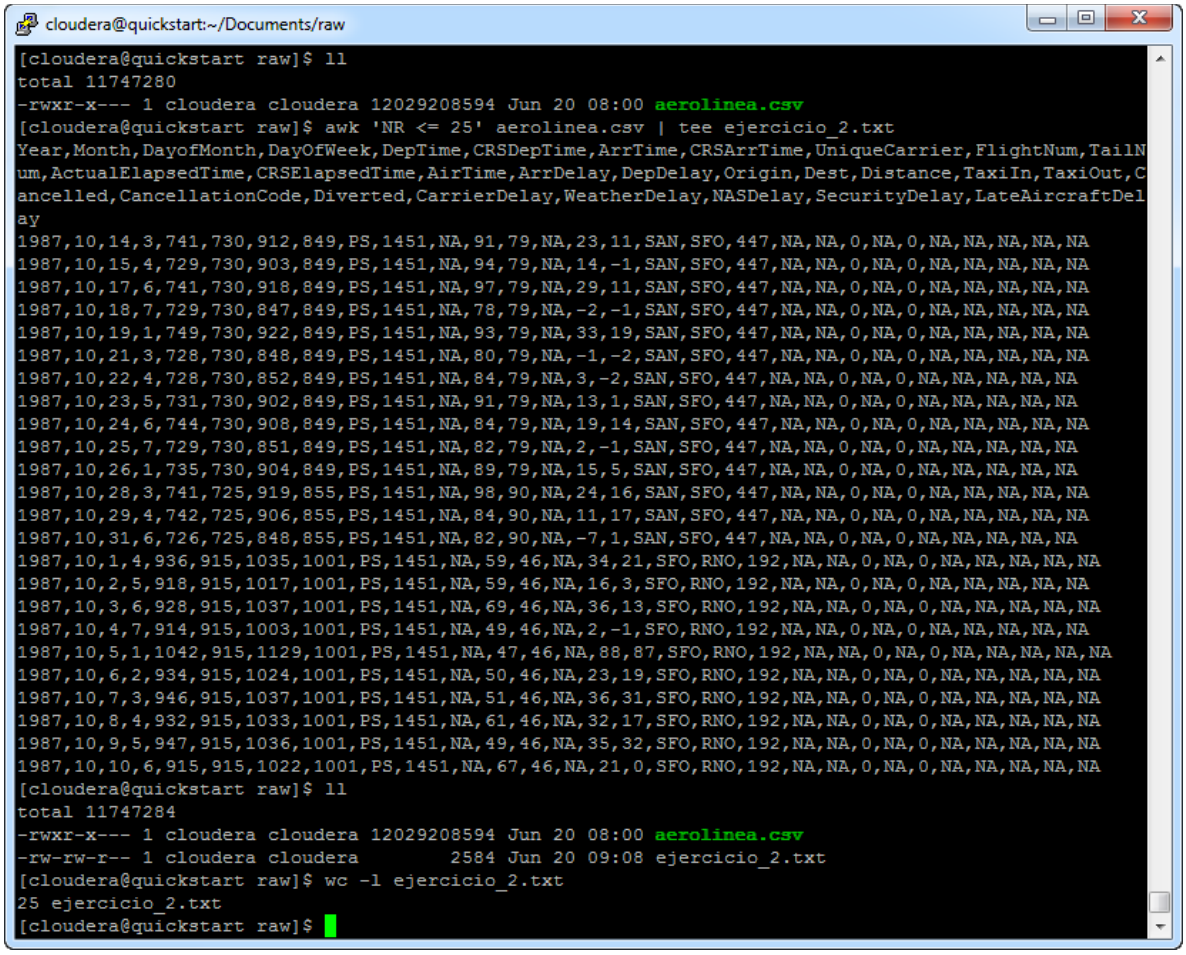
```
Respuesta: awk  'NR <=25' aerolinea.csv
```

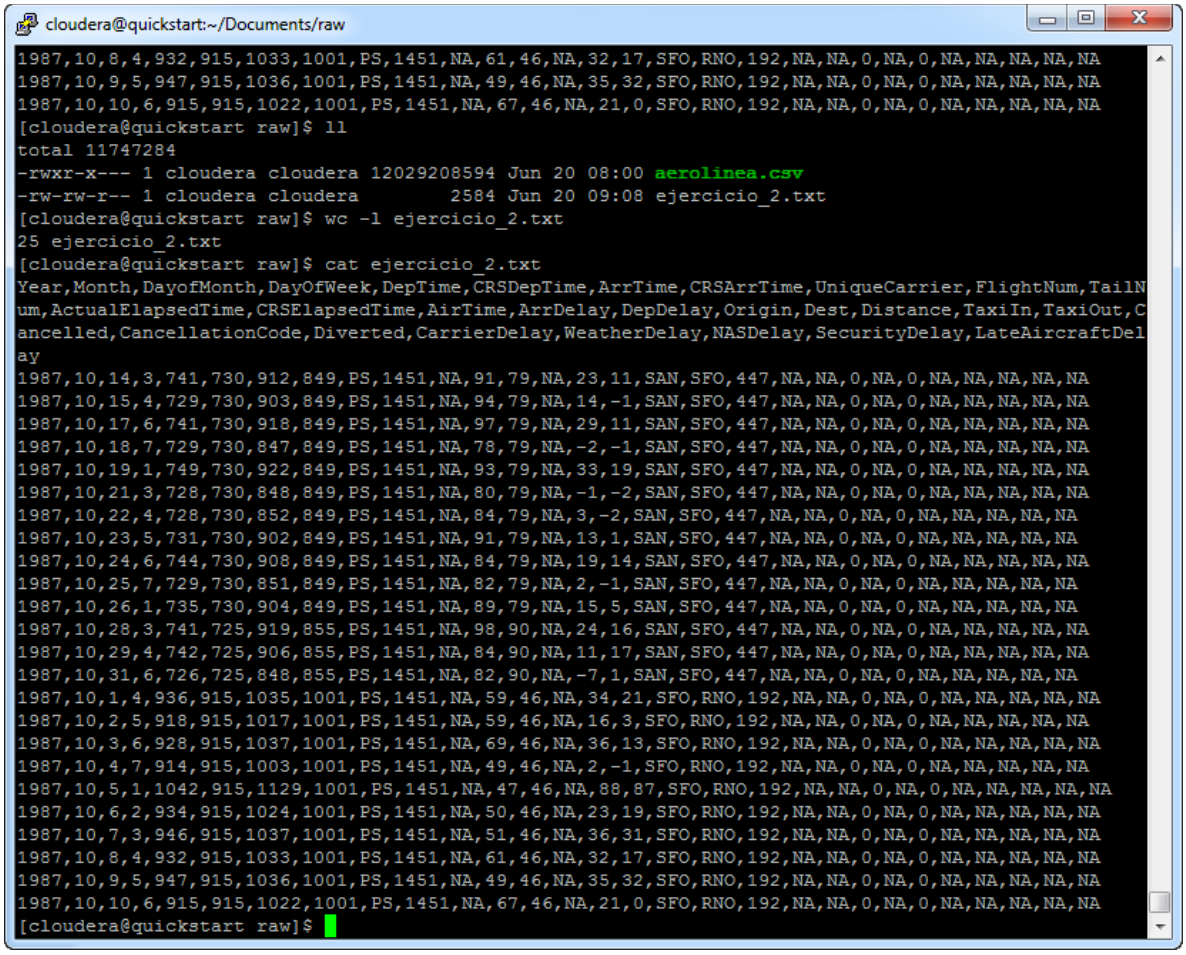


2.-Como ya se ha visto, utilizar el redireccionamiento destructivo (>) implica almacenar típicamente algún contenido en un archivo **(ej. echo "contenido" > archivo)**.
Pero lo cierto es que con este comando no se apreciará en pantalla lo que se desea almacenar en dicho archivo, por ello es que se necesita 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

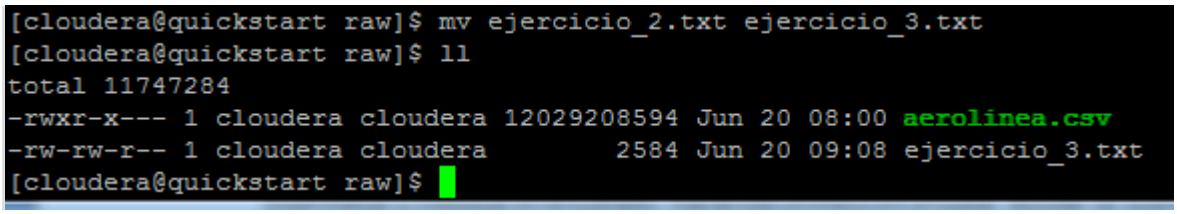
Respuesta: awk 'NR <=25' aerolinea.csv | tee ejercicio_2.txt


cloudera@quickstart:~/Documents/raw
[cloudera@quickstart raw]$ ll
total 11747280
-rwxr-x--- 1 cloudera cloudera 12029208594 Jun 20 08:00 aerolinea.csv
[cloudera@quickstart raw]$ awk 'NR <= 25' 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,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 raw]$ ll
total 11747284
-rwxr-x--- 1 cloudera cloudera 12029208594 Jun 20 08:00 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 2584 Jun 20 09:08 ejercicio_2.txt
[cloudera@quickstart raw]$ wc -l ejercicio_2.txt
25 ejercicio_2.txt
[cloudera@quickstart raw]$



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

Respuesta: mv ejercicio_2.txt ejercicio_3.txt



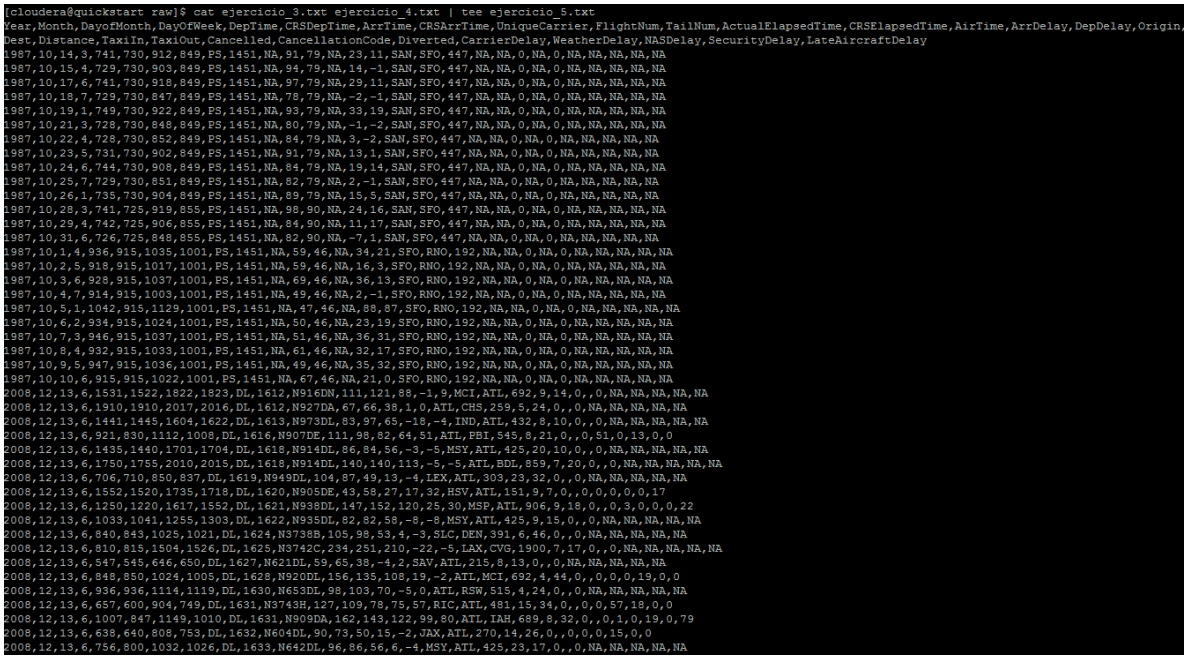
4.- Con algún comando en GNU/Linux tome las 25 últimas líneas del archivo aerolínea.csv **SIN** emplear el comando tail y guárdelo como **ejercicio_4.txt**

Respuesta: awk "NR>= $(expr $(wc -l aerolinea.csv | awk '{print $1}')-25)" aerolinea.csv | tee ejercicio_4.txt

```
[cloudera@quickstart raw]$ awk "NR>= $(expr $(wc -l aerolinea.csv | awk '{print $1}') - 25)" aeroli
nea.csv | tee ejercicio_4.txt
2008,12,13,6,1531,1522,1822,1823,DL,1612,N916DN,111,121,88,-1,9,MCI,ATL,692,9,14,0,,0,NA,NA,NA,NA,N
A
2008,12,13,6,1910,1910,2017,2016,DL,1612,N927DA,67,66,38,1,0,ATL,CHS,259,5,24,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,1441,1445,1604,1622,DL,1613,N973DL,83,97,65,-18,-4,IND,ATL,432,8,10,0,,0,NA,NA,NA,NA,N
A
2008,12,13,6,921,830,1112,1008,DL,1616,N907DE,111,98,82,64,51,ATL,PBI,545,8,21,0,,0,51,0,13,0,0
2008,12,13,6,1435,1440,1701,1704,DL,1618,N914DL,86,84,56,-3,-5,MSY,ATL,425,20,10,0,,0,NA,NA,NA,NA,N
A
2008,12,13,6,1750,1755,2010,2015,DL,1618,N914DL,140,140,113,-5,-5,ATL,BDL,859,7,20,0,,0,NA,NA,NA,NA
,NA
2008,12,13,6,706,710,850,837,DL,1619,N949DL,104,87,49,13,-4,LEX,ATL,303,23,32,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,1552,1520,1735,1718,DL,1620,N905DE,43,58,27,17,32,HSV,ATL,151,9,7,0,,0,0,0,0,0,17
2008,12,13,6,1250,1220,1617,1552,DL,1621,N938DL,147,152,120,25,30,MSP,ATL,906,9,18,0,,0,3,0,0,0,22
2008,12,13,6,1033,1041,1255,1303,DL,1622,N935DL,82,82,58,-8,-8,MSY,ATL,425,9,15,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,840,843,1025,1021,DL,1624,N3738B,105,98,53,4,-3,SLC,DEN,391,6,46,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,810,815,1504,1526,DL,1625,N3742C,234,251,210,-22,-5,LAX,CVG,1900,7,17,0,,0,NA,NA,NA,NA
,NA
2008,12,13,6,547,545,646,650,DL,1627,N621DL,59,65,38,-4,2,SAV,ATL,215,8,13,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,848,850,1024,1005,DL,1628,N920DL,156,135,108,19,-2,ATL,MCI,692,4,44,0,,0,0,0,19,0,0
2008,12,13,6,936,936,1114,1119,DL,1630,N653DL,98,103,70,-5,0,ATL,RSW,515,4,24,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,657,600,904,749,DL,1631,N3743H,127,109,78,75,57,RIC,ATL,481,15,34,0,,0,0,57,18,0,0
2008,12,13,6,1007,847,1149,1010,DL,1631,N909DA,162,143,122,99,80,ATL,IAH,689,8,32,0,,0,1,0,19,0,79
2008,12,13,6,638,640,808,753,DL,1632,N604DL,90,73,50,15,-2,JAX,ATL,270,14,26,0,,0,0,0,15,0,0
2008,12,13,6,756,800,1032,1026,DL,1633,N642DL,96,86,56,6,-4,MSY,ATL,425,23,17,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,612,615,923,907,DL,1635,N907DA,131,112,103,16,-3,GEG,SLC,546,5,23,0,,0,0,0,16,0,0
2008,12,13,6,749,750,901,859,DL,1636,N646DL,72,69,41,2,-1,SAV,ATL,215,20,11,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,1002,959,1204,1150,DL,1636,N646DL,122,111,71,14,3,ATL,IAD,533,6,45,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,834,835,1021,1023,DL,1637,N908DL,167,168,139,-2,-1,ATL,SAT,874,5,23,0,,0,NA,NA,NA,NA,N
A
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
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
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
[cloudera@quickstart raw]$
```

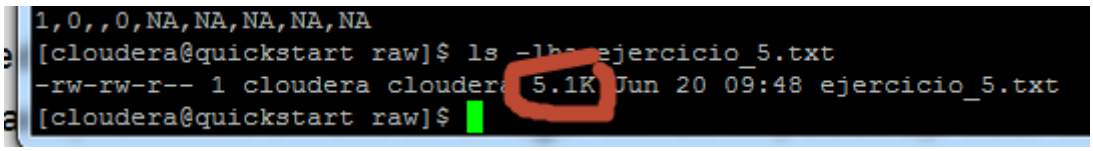
5.- Concatene los archivos **ejercicio_3.txt** y **ejercicio_4.txt** en un archivo **ejercicio_5.txt** y en esa misma pantalla resultado muestre el contenido de **ejercicio_5.txt**

Respuesta: cat ejercicio_3.txt ejercicio_4.txt | tee ejercicio_5.txt



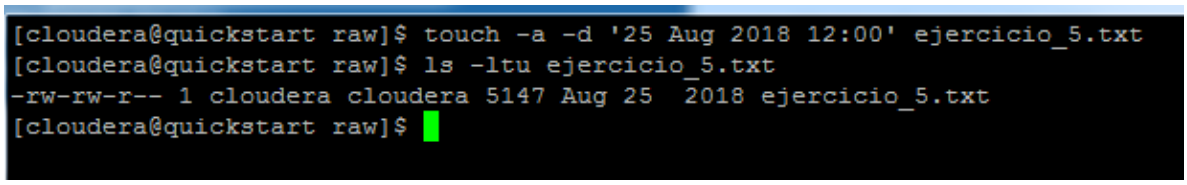
6.- Usando el comando **ls** y sus opciones, verifique el peso de **ejercicio_5.txt,** señalando en la captura de pantalla dónde se encuentra éste.

Respuesta: ls -lha ejercicio_5.txt



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.

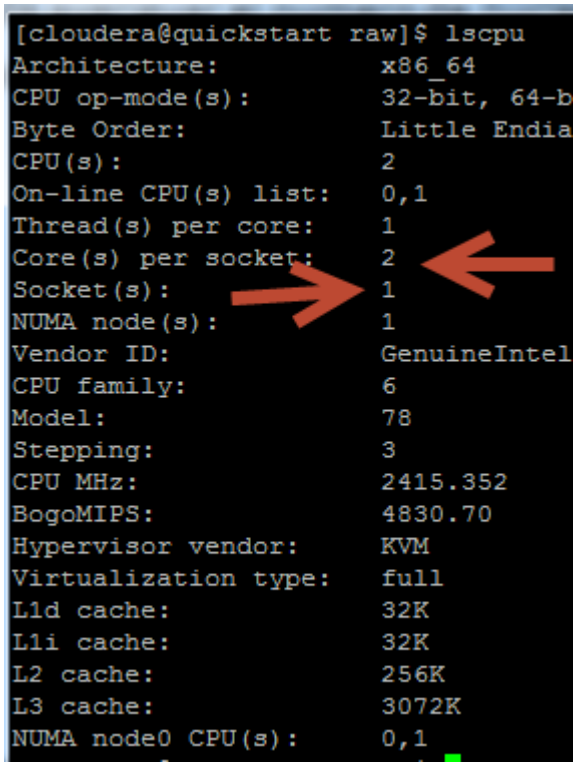
Respuesta: touch -a -d '25 Aug 2018 12:00' ejercicio_5.txt



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.

Lscpu

Core(s) per socket:
Socket (s)



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$
```

```
awk -F “,” ‘{print $3 $5}' ejercicio_5.txt
```

```
[cloudera@quickstart raw]$ 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
```

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.

```
awk -F "," '{$2="-1"; print $0}' ejercicio_5.txt > archivo_6.txt
cat archivo_6.txt
```

```
[cloudera@quickstart raw]$ awk -F "," '{$2="-1"; print $0}' ejercicio_5.txt >arc
hivo_6.txt
[cloudera@quickstart raw]$ cat archivo_6.txt
Year -1 DayofMonth DayOfWeek DepTime CRSDepTime ArrTime CRSArrTime UniqueCarrier
 FlightNum TailNum ActualElapsedTime CRSElapsedTime AirTime ArrDelay DepDelay Or
igin Dest Distance TaxiIn TaxiOut Cancelled CancellationCode Diverted CarrierDel
ay WeatherDelay NASDelay SecurityDelay LateAircraftDelay
1987 -1 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 -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 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 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 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 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 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 N
A 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 N
```

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.

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

El comando -head no forma parte de los comandos del FileSystemShell de Hadoop

Fuente:

https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-common/FileSystemShell.html

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/aerolinea.csv | wc -l
```

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

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 /raw/aerolinea.csv
```

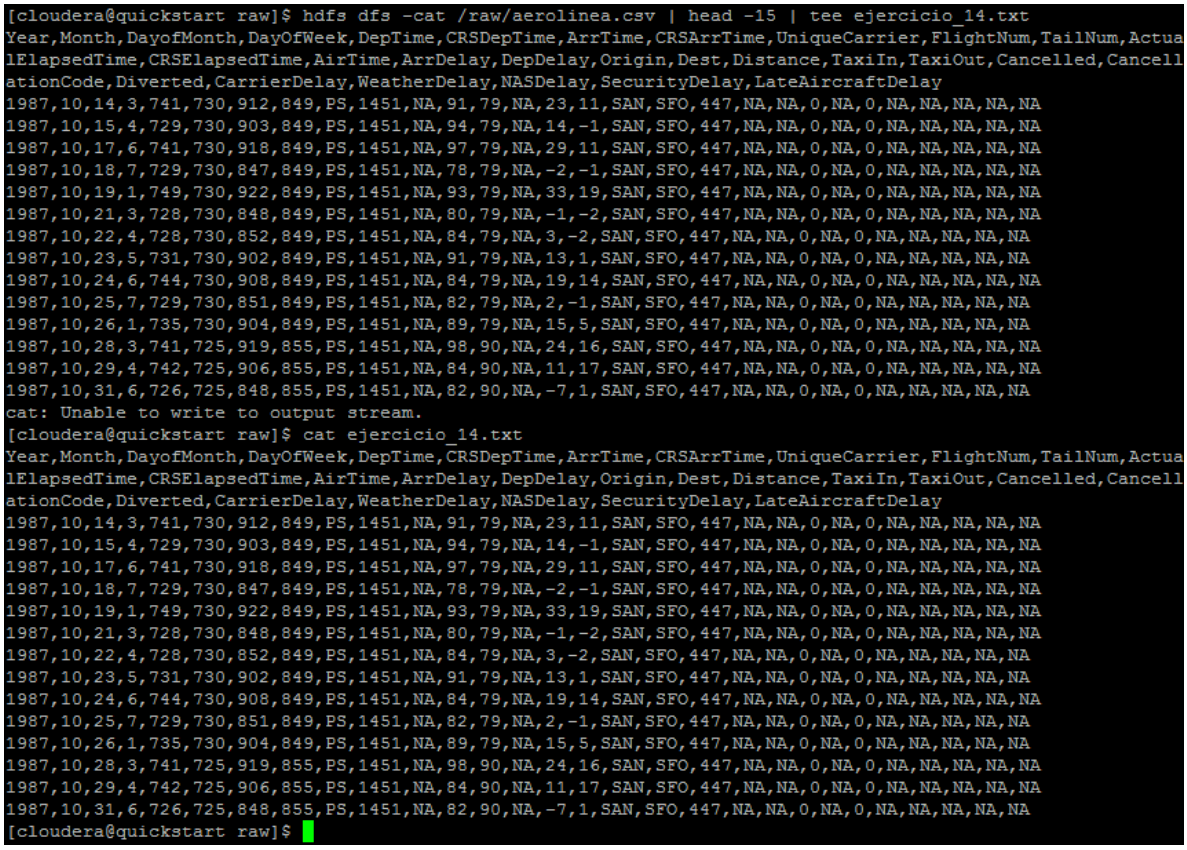
```
[cloudera@quickstart raw]$ hdfs dfs -stat %r /raw/aerolinea.csv
1
```

```
hdfs dfs -ls /raw/aerolinea.csv  (La columna después de los permisos indica el factor de réplica)
```

```
[cloudera@quickstart raw]$ hdfs dfs -ls /raw/aerolinea.csv
-rw-r--r--   1 cloudera supergroup 12029208594 2018-06-13 16:32 /raw/aerolinea.c
sv
```

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 /raw/aerolinea.csv | head -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.

Respuesta1:

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

sudo –u hdfs hdfs dfs -mkdir /stagin

sudo –u hdfs hdfs dfs -ls /master
```

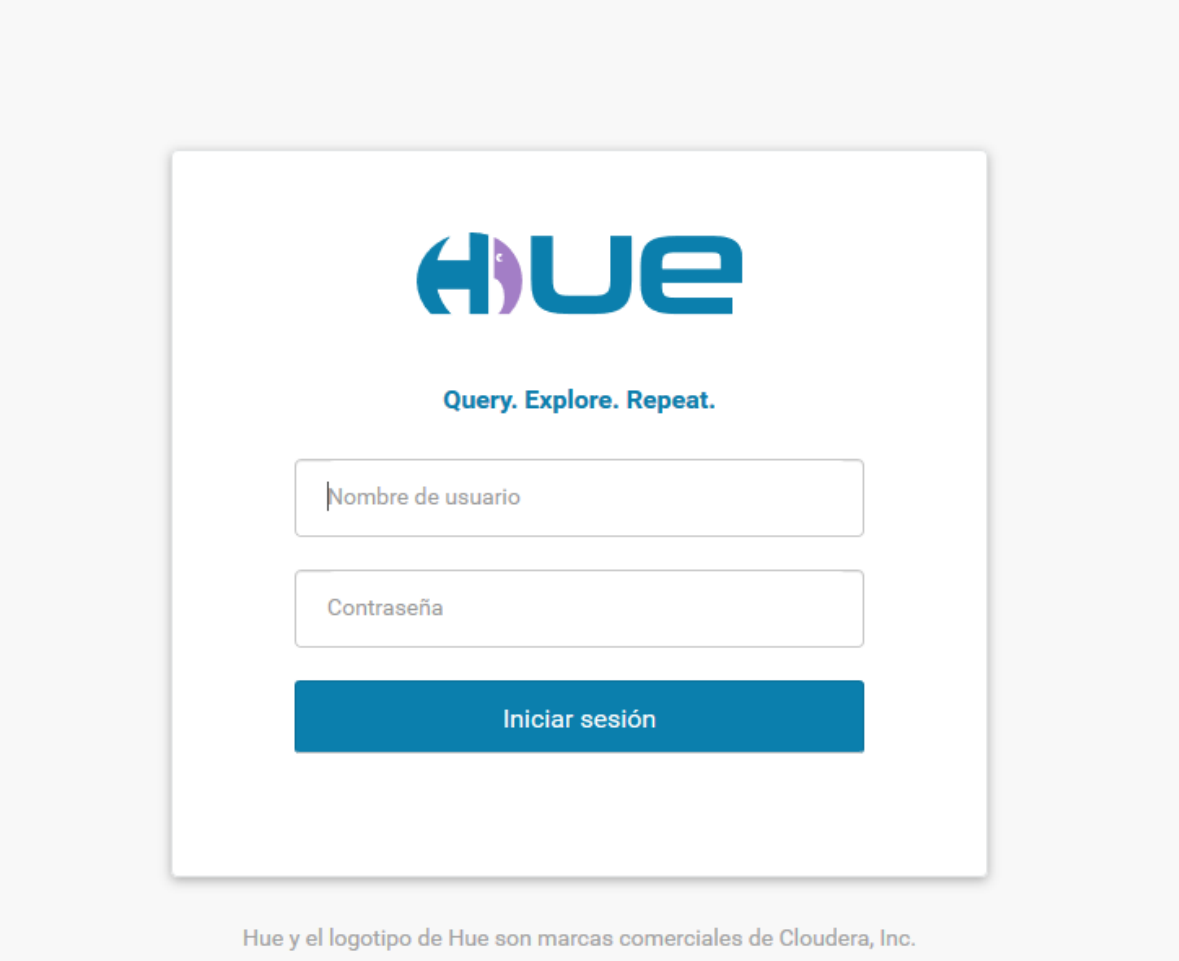
```
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -mkdir /master
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -mkdir /stagin
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -ls -lau /stagin
-ls: Illegal option -lau
Usage: hadoop fs [generic options] -ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -ls -lha /stagin
-ls: Illegal option -lha
Usage: hadoop fs [generic options] -ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -ls /stagin
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -ls /
Found 9 items
drwxrwxrwx   - hdfs     supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2018-06-25 06:24 /hbase
drwxr-xr-x   - hdfs     supergroup          0 2018-06-25 07:15 /master
drwxr-xr-x   - cloudera supergroup          0 2018-06-13 16:32 /raw
drwxr-xr-x   - solr     solr                0 2017-07-19 05:37 /solr
drwxr-xr-x   - hdfs     supergroup          0 2018-06-25 07:16 /stagin
drwxrwxrwt   - hdfs     supergroup          0 2018-06-12 16:37 /tmp
drwxr-xr-x   - hdfs     supergroup          0 2018-06-25 07:15 /user
drwxr-xr-x   - hdfs     supergroup          0 2017-07-19 05:36 /var
```

Respuesta2: sudo –u hdfs hdfs dfs –chmod 760 /raw/aerolinea.csv

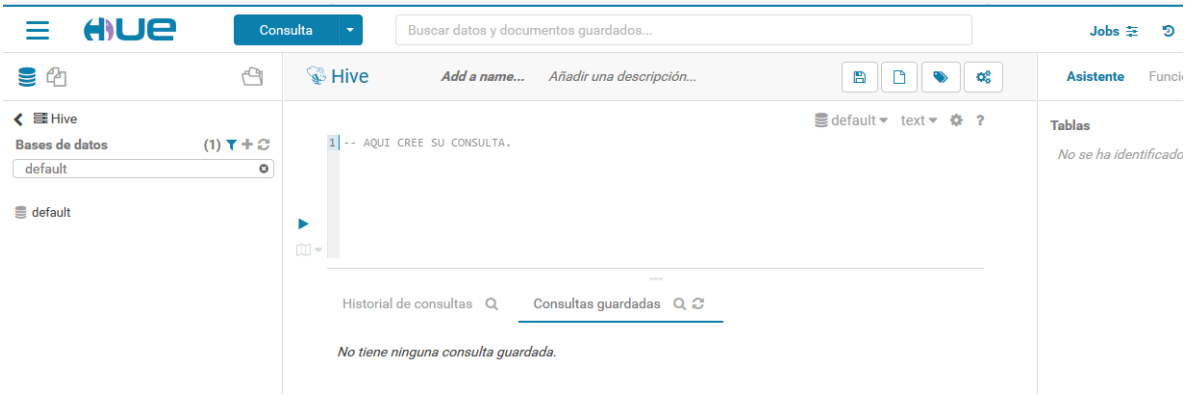
sudo –u hdfs hdfs dfs –ls /raw/aerolinea.csv

```
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -chmod 760 /raw/aerolinea.csv
[cloudera@quickstart raw]$ sudo -u hdfs hdfs dfs -ls /raw/aerolinea.csv
-rwxrw----   1 cloudera cloudera 12029208594 2018-06-13 16:32 /raw/aerolinea.csv
```

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.

```
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'
tblproperties ("skip.header.line.count"="1");
```

```
    > LateAircraftDelay STRING)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE
    > location '/raw'
    > tblproperties ("skip.header.line.count"="1");
OK
Time taken: 2.24 seconds
hive>
```

```
hive > SELECT * FROM TABLA_AEROLINEA LIMIT 10;
```

```
hive> SELECT * FROM TABLA_AEROLINEA LIMIT 10;
OK
1987    10      14      3       741     730     912     849     PS      1451    NA      91      79      NA2
3       11      SAN     SFO     447     NA      NA      0       NA      0       NA      NA      NA      NAN
A
1987    10      15      4       729     730     903     849     PS      1451    NA      94      79      NA1
4       -1      SAN     SFO     447     NA      NA      0       NA      0       NA      NA      NA      NAN
A
1987    10      17      6       741     730     918     849     PS      1451    NA      97      79      NA2
9       11      SAN     SFO     447     NA      NA      0       NA      0       NA      NA      NA      NAN
A
```

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.

```
hive> DROP TABLE TABLA_AEROLINEA;
OK
Time taken: 0.349 seconds
```

```
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 INT,
Dest STRING,
Distance STRING,
TaxiIn STRING,
TaxiOut STRING,
Cancelled STRING,
CancellationCode STRING,
Diverted STRING,
CarrierDelay STRING,
WeatherDelay STRING,
NASDelay STRING,
SecurityDelay STRING,
LateAircraftDelay STRING)
```

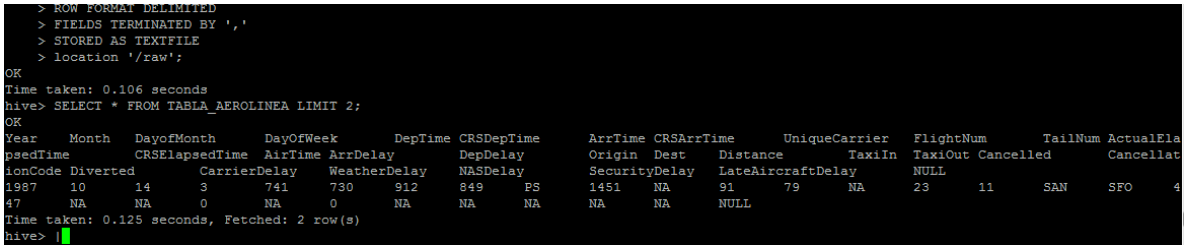
```
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
location '/raw'
tblproperties ("skip.header.line.count"="1");
```

```
hive> SELECT * FROM TABLA_AEROLINEA LIMIT 2
    > ;
OK
1987    10      14      3       741     730     912     849     PS      1451    NA      91      79      NA      23      11      NULL    SFO
47      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      NULL    SFO
47      NA      NA      0       NA      0       NA      NA      NA      NA      NA
Time taken: 0.118 seconds, Fetched: 2 row(s)
```

Respuesta: Las observaciones tienen la variable Origin como NULL dado que no corresponde el tipo de dato y no hace un casteo en automático.

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.

```
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';
```



La tabla se crea con una columna adicional en NULL dado que no encontró datos con esa descripción en el archivo fuente.

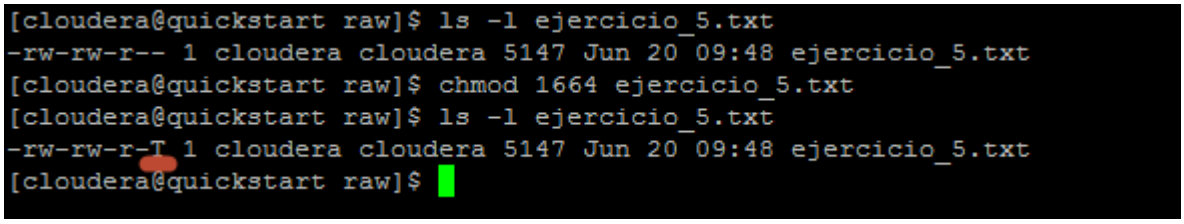
19.- En esta tabla anterior inserte un renglón a la tabla con todos los valores iguales a "NA" (tiene que investigar cómo añadir elementos a la tabla), y luego después de la inserción del elemento indague en qué parte del HDFS se ha guardado ese nuevo elemento.

```
INSERT INTO TABLA_AEROLINEA VALUES ("NA", "NA", "NA", "NA", "NA", "NA", "NA",
"NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA",
"NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA")
```

SECCIÓN 3. PREGUNTAS ABIERTAS

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

Sticky Bit es una bandera de acceso para el propietario de un archivo, con ese bit el sistema de archivos solamente permite que el propietario o el usuario root puedan renombrar o borrar el archivo.



Se indica con una "t" o "T" en el lugar del permiso de ejecución de "OTROS"

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

NoSQL se le llama a las bases de datos no relacionales que sirven para modelos de datos sin esquema y escalables.

Hive es SQL dado que las bases de datos son estructuradas, adicional a que hive es una interface que permite ejecutar consultas y transformarlas a jobs de map reduce con la finalidad de operar en sistemas de archivos distribuidos por medio de un framework de batches.

Impala es una base de datos analítica para hadoop que ejecuta las consultas de SQL en hadoop, pero sus sistema primario de base de datos es relacional DBMS, por lo cual no es un sistema NoSQL.

Fuentes:

https://www.cloudera.com/documentation/enterprise/5-8-x/topics/impala_databases.html

https://hive.apache.org/

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?

El comando nohup permite que un job se siga ejecutando en background aún cuando se cierra la sesión que lo ejecutó. La forma de utilizarlo es:

nohup <command> <options> &

Se puede utilizar en un sistema distribuido ya que se pueden lanzar diferentes comandos y jobs tipo daemon y seguir trabajando, esto es conveniente ya que ciertos jobs pueden tomar mucho tiempo y no es conveniente que puedan ser interrumpidos si la sesión tiene algún problema.

Fuente:

https://linux.101hacks.com/unix/nohup-command/

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.     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.

Los valores adecuados para saber cuanta memoria tenemos disponible para utilizar es 3.2G; esto dado que se considera la mamorio que está completamente sin utilizar y la memoria que se llegó a utilizar pero que ahora está disponible aún cuando no está vacía, mientras el valor de Memoria->libre indica 1.2G que se refiere únicamente a la memoria vacía.

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**?

La estrategia sería colocar al usuario_nuestro como owner del archivo y darle únicamente permisos de lectura, con lo cual no se afectan los permisos del grupo mediante los comandos:

```
chown usuario_nuestro:grupo_nuestro objetivo.txt

chmod 477 objetivo.txt
```

De ser posible se recomienda mover de grupo al usuario_nuestro y cambiar los permisos para OTROS usuarios con los comandos:

```
chgrp otro objetivo.txt

chmod 774 objetivo.txt
```

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

Cludera es una plataforma que integra los diferentes sistemas de un ambiente big data pre configurado y con herramientas propias para la gestión, mantenimiento y escalabilidad, dentro del ecosistema Cloudera se encuentra Hadoop. Hadoop es un framework de software que soporta aplicaciones de ámbito distribuido y bajo licencia de uso libre para procesar grandes volúmenes de datos a través de clases de servidores básicos.

Fuentes:

https://www.ibm.com/analytics/mx/es/technology/hadoop/index.html

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

```
En Linux se pueden dividir en tres tipos:
Regulares (se indican con – )
Directorios (se indican con d)
Especiales: En este caso se engloban:
       -Block File (b)
       -Character Device File (c)
       - Named pipe file (p)
       - Symbolic link file (l)
       - Socket file (s)
                + '-' Regular file
                + 'd' Directory
                + 'l' Symbolic link
                + 'p' Named pipe: Connect the output or one process to the inpu$
                + 's' Socket: Used for inter-process communication.
                + 'c' / 'b' Device file.
                + 'D' Door: File communication between a client and a server.
```

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

SerDe se compone de las palabras Serialización y Deserialización, en general se refiere a un framework para poder serializar y deserializar estructuras de datos eficientemente y de manera general.

En HIVE e Impala por ejemplo SerDe se utiliza para convertir los registros en objetos de java y por tanto que Hive pueda utilizarlos, cuando serializa convierte el objeto de Java en un formato que pueda ser guardado en formato HDFS.

Fuentes:

https://serde.rs/

https://www.quora.com/What-is-SerDe-in-Hive

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

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

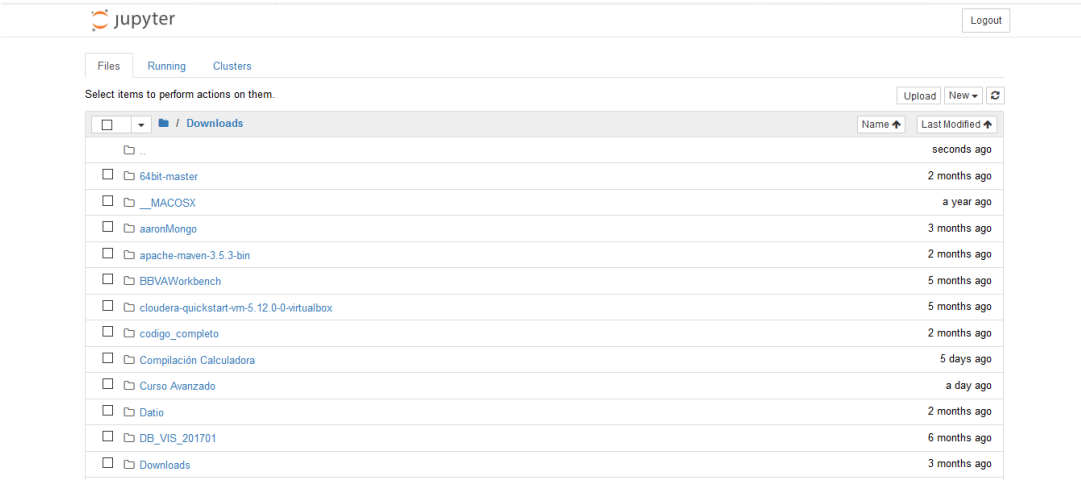
30.- ¿Existe algún otro tipo de sistemas de archivos distribuidos que NO sea HDFS? si es así, ¿de cuáles se trata?

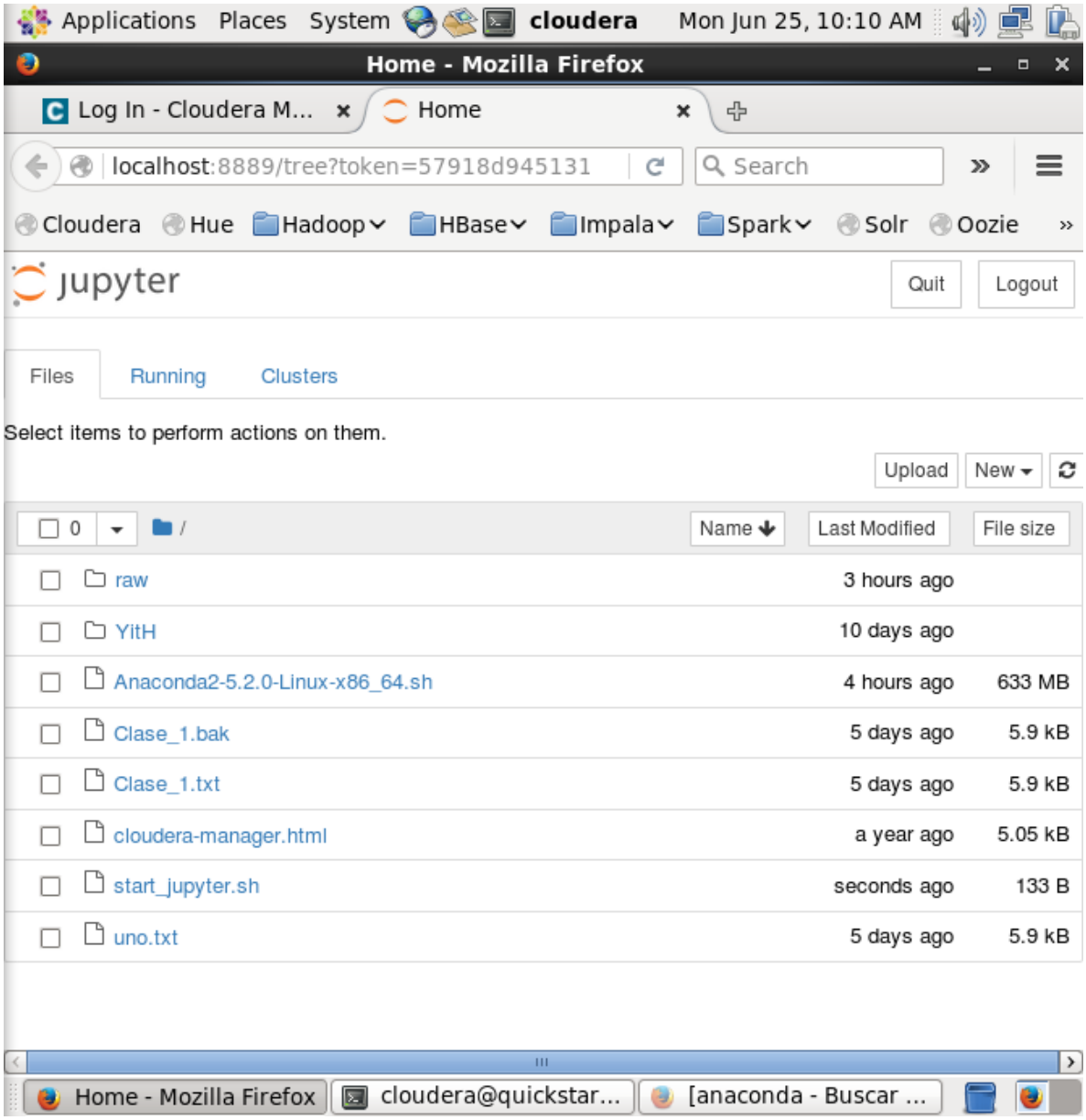
SECCIÓN 4. ESPECIAL

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:




Applications Places System cloudera Mon Jun 25, 10:10 AM
Home - Mozilla Firefox
Log In - Cloudera M...
Home
localhost:8889/tree?token=57918d945131
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie
jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
0
/
Name
Last Modified
File size
raw 3 hours ago
YitH 10 days ago
Anaconda2-5.2.0-Linux-x86_64.sh 4 hours ago 633 MB
Clase_1.bak 5 days ago 5.9 kB
Clase_1.txt 5 days ago 5.9 kB
cloudera-manager.html a year ago 5.05 kB
start_jupyter.sh seconds ago 133 B
uno.txt 5 days ago 5.9 kB
Home - Mozilla Firefox
cloudera@quickstar...
[anaconda - Buscar ...