

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 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 ' aerolíneas.csv`

```
^C
[cloudera@quickstart compartida]$ awk 'FNR <= 25' aerolineas.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
```

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 ' aerolíneas.csv | tee ejercicio_2.txt`

```
116 117 118
[cloudera@quickstart compartida]$ awk 'NR <= 25' aerolineas.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,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
```

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**

```
[cloudera@quickstart compartida]$ mv ejercicio_2.txt ejercicio_3.txt
[cloudera@quickstart compartida]$ ls
aerolineas.csv  ejercicio_3.txt
[cloudera@quickstart compartida]$
```

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

Respuesta: **tac aerolineas.csv | head -25 | tac**

```
[cloudera@quickstart compartida]$ tac aerolineas.csv | head -25 | tac
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,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,NA,NA
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,NA,NA
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,NA
2008,12,13,6,706,710,850,837,DL,1619,N949DN,104,87,49,13,-4,IFX,ATI,303,23,32,0,0,NA,NA,NA,NA,NA,NA
```

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 compartida]$ 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,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,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
1987,10,22,4,728,730,850,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,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 compartida]$ ls -l -h
total 12G
-rwxrwxrwx 1 root root 12G Jun 12 16:28 aerolineas.csv
-rw-rw-r-- 1 cloudera cloudera 4.9K Jun 24 14:38 ejemplo.txt
-rw-rw-r-- 1 cloudera cloudera 0 Jun 24 14:18 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 24 14:12 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2.5K Jun 24 14:49 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera 5.0K Jun 24 15:32 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 -d '2018-08-25' ejercicio_5.txt`

```
[cloudera@quickstart compartida]$ ls -l
total 11747304
-rwxrwxrwx 1 root      root      12029208594 Jun 12 16:28 aerolineas.csv
-rw-rw-r-- 1 cloudera cloudera    5003 Jun 24 14:38 ejemplo.txt
-rw-rw-r-- 1 cloudera cloudera      0 Jun 24 14:18 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera   2584 Jun 24 14:12 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera   2462 Jun 24 14:49 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera   5046 Jun 24 16:01 ejercicio_5.txt
[cloudera@quickstart compartida]$ touch -d '2018-08-25' ejercicio_5.txt
[cloudera@quickstart compartida]$ ls -l
total 11747304
-rwxrwxrwx 1 root      root      12029208594 Jun 12 16:28 aerolineas.csv
-rw-rw-r-- 1 cloudera cloudera    5003 Jun 24 14:38 ejemplo.txt
-rw-rw-r-- 1 cloudera cloudera      0 Jun 24 14:18 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera   2584 Jun 24 14:12 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera   2462 Jun 24 14:49 ejercicio_4.txt
-rw-rw-r-- 1 cloudera cloudera   5046 Aug 25  2018 ejercicio_5.txt
[cloudera@quickstart compartida]$
```

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`

```
[cloudera@quickstart ~]$ nproc
2
```

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:

Respuesta: `awk -F ' ' '{print $3 " " $5}' ejercicio_5.txt`

```
[cloudera@quickstart compartida]$ 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|728
```

```
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$
```

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: `awk -F ',' '{ $2 = "-1"; print }' ejercicio_5.txt > ejercicio_6.txt`

```
[cloudera@quickstart compartida]$ awk -F ',' '{ $2 = "-1"; print }' ejercicio_5.txt > ejercicio_6.txt
[cloudera@quickstart compartida]$ cat ejercicio_6.txt
Year -1 DayofMonth DayofWeek DepTime CRSDepTime ArrTime CRSArrTime UniqueCarrier FlightNum TailNum ActualElapsedTime CRSElapsedTime AirTime ArrDelay DepDelay Origin Des
t Distance TaxiIn TaxiOut Cancelled CancellationCode Diverted CarrierDelay 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
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
```

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:

```
[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 /prueba/aerolíneas.csv | wc -l`

```
[cloudera@quickstart compartida]$ hdfs dfs -cat /prueba/aerolineas.csv | wc -l
123534972
[cloudera@quickstart compartida]$
```

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 -du -h /prueba/aerolíneas.csv`

Muestra factor de replica en porcentaje

```
[cloudera@quickstart ~]$ hdfs dfs -stat %r /prueba/aerolineas.csv
1
```

Muestra factor de replica en peso

```
[cloudera@quickstart ~]$ hdfs dfs -du -h /prueba/aerolineas.csv
11.2 G 11.2 G /prueba/aerolineas.csv
```

Peso original

Peso de replica

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 /prueba/aerolíneas.csv | head -15 | tee ejercicio_14.txt`

```
[cloudera@quickstart ~]$ hdfs dfs -cat /raw/aerolineas.csv | head -15 | tee ejercicio_14.txt
Year,Month,DayOfMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailN
Dest,Distance,TaxiIn,TaxiOut,Canceled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay
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 ~]$ ls
archivo.csv  cm api.py  derby.log  Documents  eclipse     enterprise-deployment.json  k
cloudera-manager  Desktop    Downloads  ejercicio_14.txt  express-deployment.json    l
```

15.- Cree los directorios **master** y **staging** 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`

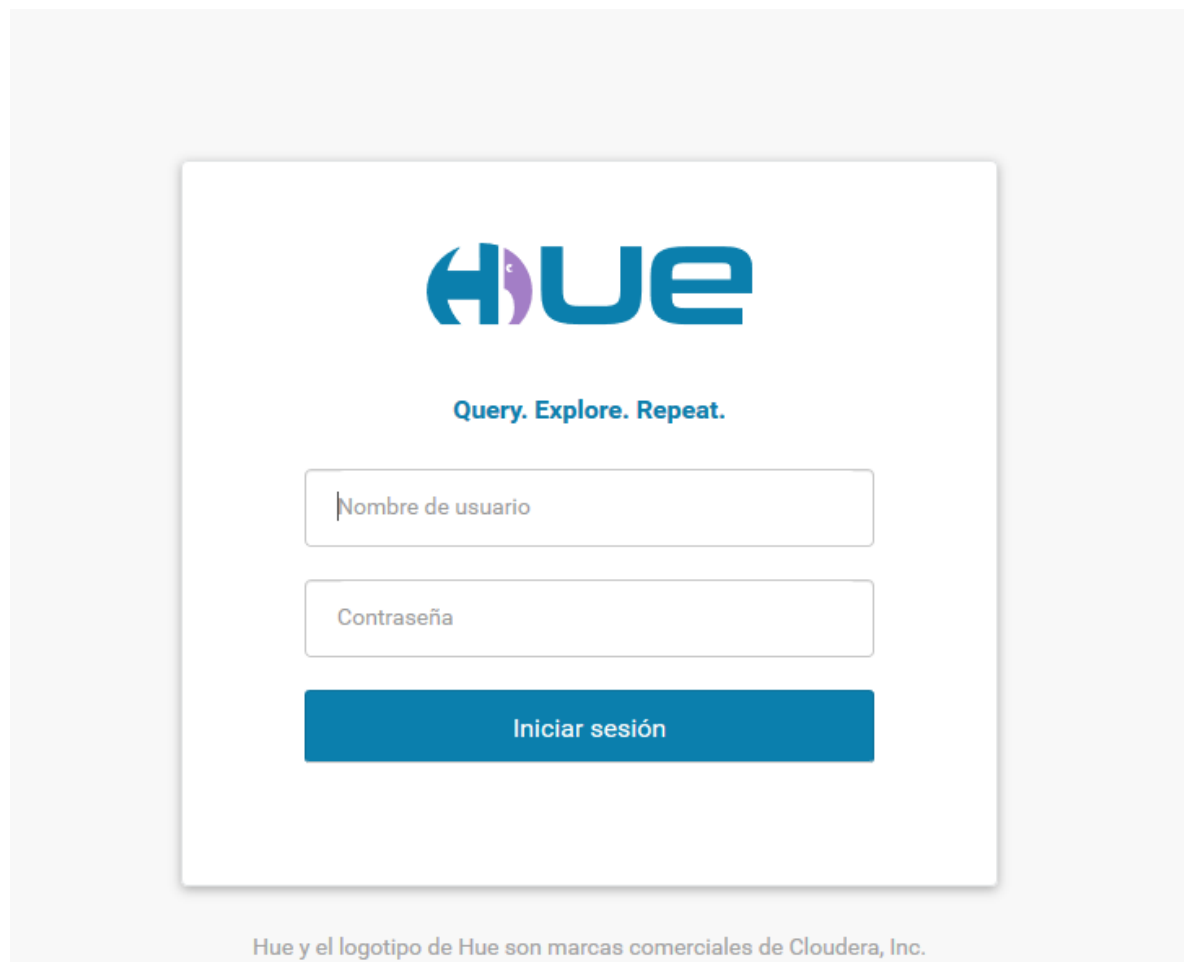
`sudo -u hdfs hdfs dfs -mkdir /master`

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

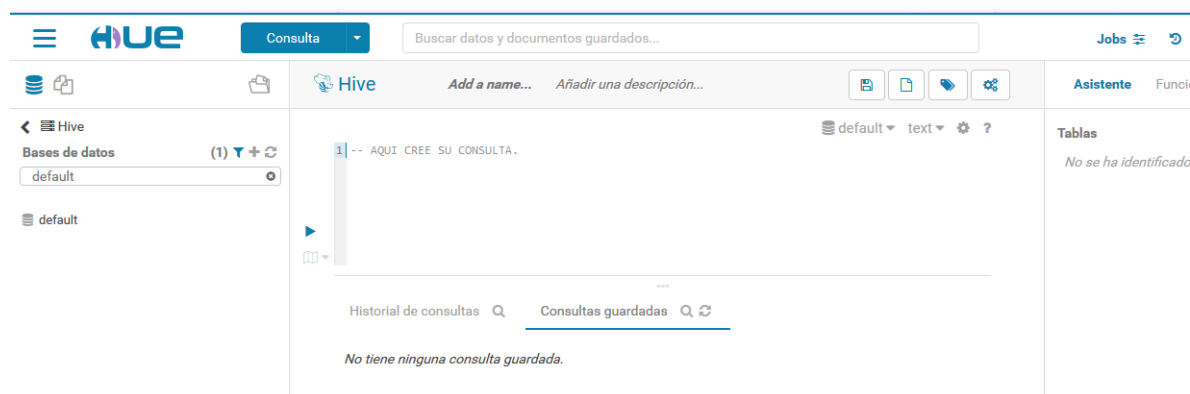
```
[cloudera@quickstart compartida]$ sudo -u hdfs hdfs dfs -mkdir /master
[cloudera@quickstart compartida]$ sudo -u hdfs hdfs dfs -mkdir /staging
[cloudera@quickstart compartida]$ sudo -u hdfs hdfs dfs -chmod 760 /raw/aerolinea.csv
[cloudera@quickstart compartida]$ hdfs dfs -ls /raw/
Found 3 items
-rwxrw---- 1 cloudera supergroup 12029208594 2018-06-13 16:34 /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,
```


Garcia Cruz Mario Alberto

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')`

Hive `-e "select * from tabla_aerolinea limit 10"`

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

```
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
Time taken: 2.156 seconds
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
```

```
[cloudera@quickstart ~]$ hive -e "select * from tabla_aerolinea limit 10"
```

```
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
1987 10 14 3 741 730 912 849 PS 1451 NA 91 79 NA 23 11 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 94 79 NA 14 -1 SAN SFO 447 NA NA
1987 10 15 4 729 730 903 849 PS 1451 NA 94 79 NA 14 -1 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 97 79 NA 29 11 SAN SFO 447 NA NA
1987 10 17 6 741 730 918 849 PS 1451 NA 97 79 NA 29 11 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 78 79 NA -2 -1 SAN SFO 447 NA NA
1987 10 18 7 729 730 847 849 PS 1451 NA 78 79 NA -2 -1 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 93 79 NA 33 19 SAN SFO 447 NA NA
1987 10 19 1 749 730 922 849 PS 1451 NA 93 79 NA 33 19 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 80 79 NA -1 -2 SAN SFO 447 NA NA
1987 10 21 3 728 730 848 849 PS 1451 NA 80 79 NA -1 -2 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 84 79 NA 3 -2 SAN SFO 447 NA NA
1987 10 22 4 728 730 852 849 PS 1451 NA 84 79 NA 3 -2 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 91 79 NA 13 1 SAN SFO 447 NA NA
1987 10 23 5 731 730 902 849 PS 1451 NA 91 79 NA 13 1 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 84 79 NA 19 14 SAN SFO 447 NA NA
1987 10 24 6 744 730 908 849 PS 1451 NA 84 79 NA 19 14 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 82 79 NA 2 -1 SAN SFO 447 NA NA
1987 10 25 7 729 730 851 849 PS 1451 NA 82 79 NA 2 -1 SAN SFO 447 NA NA
NA 0 NA NA NA NA NA NA
```

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

```
[cloudera@quickstart ~]$ hive -e "select * from tabla_aerolinea limit 10"
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties
OK
1987 10 14 3 741 730 912 849 PS 1451 NA 91 79 NA 23 11 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 94 79 NA 14 -1 NULL SFO 447 NA NA 0
1987 10 15 4 729 730 903 849 PS 1451 NA 94 79 NA 14 -1 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 97 79 NA 29 11 NULL SFO 447 NA NA 0
1987 10 17 6 741 730 918 849 PS 1451 NA 97 79 NA 29 11 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 78 79 NA -2 -1 NULL SFO 447 NA NA 0
1987 10 18 7 729 730 847 849 PS 1451 NA 78 79 NA -2 -1 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 93 79 NA 33 19 NULL SFO 447 NA NA 0
1987 10 19 1 749 730 922 849 PS 1451 NA 93 79 NA 33 19 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 80 79 NA -1 -2 NULL SFO 447 NA NA 0
1987 10 21 3 728 730 848 849 PS 1451 NA 80 79 NA -1 -2 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 84 79 NA 3 -2 NULL SFO 447 NA NA 0
1987 10 22 4 728 730 852 849 PS 1451 NA 84 79 NA 3 -2 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 91 79 NA 13 1 NULL SFO 447 NA NA 0
1987 10 23 5 731 730 902 849 PS 1451 NA 91 79 NA 13 1 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 84 79 NA 19 14 NULL SFO 447 NA NA 0
1987 10 24 6 744 730 908 849 PS 1451 NA 84 79 NA 19 14 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA 849 PS 1451 NA 82 79 NA 2 -1 NULL SFO 447 NA NA 0
1987 10 25 7 729 730 851 849 PS 1451 NA 82 79 NA 2 -1 NULL SFO 447 NA NA 0
NA 0 NA NA NA NA NA NA
```

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.

```
[cloudera@quickstart ~]$ hive -e "select * from tabla_aerolinea limit 10"
```

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.12.0.jar!/hive-log4j.properties

OK																					
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	NULL														
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	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.

Respuesta: `chmod o+t ejercicio_5.txt`

```
[cloudera@quickstart compartida]$ chmod o+t ejercicio_5.txt
[cloudera@quickstart compartida]$ ls -l
total 11747316
-rwxrwxrwx 1 root root 12029208594 Jun 12 16:28 aerolineas.csv
-rw-rw-r-- 1 cloudera cloudera 5003 Jun 24 14:38 ejemplo.txt
-rw-rw-r-- 1 cloudera cloudera 0 Jun 24 14:18 ejercicio_2.txt
-rw-rw-r-- 1 cloudera cloudera 2584 Jun 24 14:12 ejercicio_3.txt
-rw-rw-r-- 1 cloudera cloudera 2462 Jun 24 14:49 ejercicio_4.txt
-rw-rw-r-T 1 cloudera cloudera 5046 Jun 24 17:00 ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5043 Jun 24 17:04 ejercicio_6.txt
-rw-rw-r-- 1 cloudera cloudera 0 Jun 24 16:42 ejercicio6.txt
```

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

Respuesta: NoSQL se refiere a la no relacion de estructura y tipos de datos permitiendo el manejo de grandes cantidades de datos, perdiendo integridad en la información.

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: el comando `nohup` previene que el proceso en curso se aborte cuando el usuario cierre sesión o se desconecte, en un sistema distribuido permite ejecutar varios procesos sin la necesidad de que termine uno para ejecutar los demás.

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.

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: `chmod 755 objetivo.txt`

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

Respuesta: Cloudera es un conjunto de herramientas que proporcionan un entorno completo de Big Data incluyendo Hadoop, mientras que Hadoop es una herramienta para optimizar el procesamiento de un entorno BigData

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

Respuesta: Archivos físicos, Directorios, Enlaces, Enlaces simbólicos

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

Respuesta: Indica la manera en que se debe procesar un archivo de Hive o Impala, es una combinación de Serializer y Deserializer. Donde:

- **Deserializer:** toma un string o binario y lo convierte en objeto java
- **Serializer:** toma un objeto Java y lo convierte en algo que pueda escribir a HDFS

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

- **Big Table:** es un sistema de Google distribuido que almacena y procesa grandes volúmenes de archivos.
- **Big Query:** es un servicio web de Google que permite consultar grandes volúmenes de datos

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

- **Data Lake :** Se utiliza para la gestión de ingesta de datos, su almacenamiento y procesamiento, para brindar acceso a los resultados, esta soportado por un sistema de archivos.
- **Data Warehouse:** permite el proceso de información de diferentes fuentes en un ambiente estructurado.

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


- **CODA**
- **GLUSTERFS**
- **OPEN GFS**

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

/ Downloads

NameLast 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