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.

---> sed –n 1,25p aerolínea.csv

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

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

----> sed –n 1,25p aerolínea.csv | tee archivo_2.txt

```
[cloudera@quickstart Desktop]$ sed -n 1,25p aerolinea.csv | tee archivo_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 Desktop]$
```

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

-----> mv archivo_2.txt archivo_3.txt

```
[cloudera@quickstart Desktop]$ mv archivo_2.txt archivo_3.txt
[cloudera@quickstart Desktop]$ cat archivo_3.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 Desktop]$
```

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**

------> tac aerolínea.csv|head -25|tac >archivo_4.txt

```
[cloudera@quickstart Desktop]$ tac aerolinea.csv|head -25|tac > archivo_4.txt
[cloudera@quickstart Desktop]$ cat archivo_4.txt
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,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
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,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
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 Desktop]$
```

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**

-----> cat archivo_3.txt archivo_4.txt|tee archivo_5.txt

```
[cloudera@quickstart Desktop]$ cat archivo_3.txt archivo_4.txt|tee archivo_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
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
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,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
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,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
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
```

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.

-------> ls –l –h archivo_5.txt

```
[cloudera@quickstart Desktop]$ ls -l -h archivo_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.0K Jun 24 16:23 archivo_5.txt
[cloudera@quickstart Desktop]$
```

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.

-----> touch –d “2018-08-25” archivo_5.txt

```
[cloudera@quickstart Desktop]$ touch -d "2018-08-25" archivo_5.txt
[cloudera@quickstart Desktop]$ ls -l -h archivo_5.txt
-rw-rw-r-- 1 cloudera cloudera 5.0K Aug 25  2018 archivo_5.txt
[cloudera@quickstart Desktop]$
```

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.

----> cat /proc/cpuinfo

```
[cloudera@quickstart Desktop]$ cat /proc/cpuinfo
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 78
model name      : Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz
stepping        : 3
cpu MHz         : 2592.000
cache size      : 4096 KB
physical id     : 0
siblings        : 1
core id         : 0
cpu cores       : 1
apicid          : 0
initial apicid  : 0
fpu             : yes
fpu_exception   : yes
cpuid level     : 22
wp              : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush mmx fxsr sse sse2 syscall nx rdtscp lm constant_tsc up rep_goo
d xtopology nonstop_tsc unfair_spinlock pni pclmulqdq monitor ssse3 cx16 sse4_1 sse4_2 movbe popcnt aes xsave avx rdrand hypervisor lahf_lm abm 3dnowprefetch av
x2 rdseed
bogomips        : 5184.00
clflush size    : 64
cache_alignment : 64
address sizes   : 39 bits physical, 48 bits virtual
power management:
```

```
[cloudera@quickstart Desktop]$ dir /proc
1      17001  23   343   4541  5     5895  7549  8204  8330  8461  8592  9376  9446  9525  9693       dma          kmsg          partitions     tty
10     17007  24   344   4553  5040  6     756   8215  8332  8469  8595  9377  9448  9548  9694       driver       kpagecount    sched_debug    uptime
11     17237  25   35    4565  5076  6173  757   8219  8338  8471  8596  9389  9457  9641  9695       execdomains  kpageflags    schedstat      version
12     18     26   36    4566  5169  6571  7617  8224  8339  8480  8664  9390  9458  9645  9707       fb           loadavg       scsi           vmallocinfo
13     19     27   37    46    5186  6585  7631  8230  8385  8484  888   9396  9459  9669  9722       filesystems  locks         self           vmstat
1389   191    275  4     4603  5208  6702  7669  8233  8405  8493  9     9398  9463  9670  acpi       fs           mdstat        slabinfo       zoneinfo
14     193    277  436   4612  5289  6716  7896  8236  8427  8494  9269  9405  9467  9671  buddyinfo  interrupts   meminfo       softirqs
14465  19609  28   4391  4684  5317  6827  8     8247  8428  8495  9273  9406  9468  9674  bus        iomem        misc          stat
14467  2      29   44    4694  5418  6894  8093  8292  8438  8497  9297  9419  9469  9677  cgroups    ioports      modules       swaps
14468  20     3    4413  47    5504  7     81    8300  8446  8501  9338  9425  9485  9678  cmdline    irq          mounts        sys
15     21     30   4470  48    563   716   8117  8308  8453  8505  9348  9431  9487  9680  cpuinfo    kallsyms     mtd           sysrq-trigger
16     21057  32   4493  4865  5637  7265  8159  8309  8455  8513  9359  9435  9488  9685  crypto     kcore        mtrr          sysvipc
16699  22     33   4506  4879  5761  727   8195  8324  8459  8524  9367  9439  9497  9690  devices    keys         net           timer_list
17     22557  34   4511  4920  5813  7289  82    8326  8460  8542  9368  9445  9500  9692  diskstats  key-users    pagetypeinfo  timer_stats
[cloudera@quickstart Desktop]$
```

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}’ archivo_5.txt

```
[cloudera@quickstart Desktop]$ awk -F ',' '{print $3,$5}' archivo_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 1910
13 1441
13 921
13 1435
13 1750
```

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}’ archivo_5.txt > archivo_6.txt

```
[cloudera@quickstart Desktop]$ awk -F ',' '{$2="-1";print}' archivo_5.txt > archivo_6.txt
[cloudera@quickstart Desktop]$ archivo_6.txt
bash: archivo_6.txt: command not found
[cloudera@quickstart Desktop]$ 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 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 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 NA 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 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 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 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 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 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 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 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 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 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 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 NA
1987 -1 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 -1 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 -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 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 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 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 NA
2008 -1 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 -1 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
2008 -1 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 -1 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
2008 -1 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 -1 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 -1 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 -1 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
```

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 ~]$ hdfs dfs -head /raw/aerolinea.csv
-head: Unknown command
[cloudera@quickstart ~]$ hdfs dfs -tail /raw/aerolinea.csv
tail: `/raw/aerolinea.csv': No such file or directory
```

HFDS no soporta el comando head, de hecho, varios comandos simples como el ls no lo soporta de igual manera, para saber qué comandos se pueden utilizar dentro del HDFS se tiene que escribir el comando hdfs dfs.

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

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

```
[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.

----> hdfs fsck /raw/aerolínea.csv –files –blocks -racks

```
Status: HEALTHY
 Total size:    12029208594 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      90 (avg. block size 133657873 B)
 Minimally replicated blocks:   90 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Mon Jun 25 10:19:08 PDT 2018 in 10 milliseconds


The filesystem under path '/raw/aerolinea.csv' is HEALTHY
[cloudera@quickstart ~]$ hdfs fsck /raw/aerolinea.csv -files -blocks -racks
```

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/caerolinea.csv | sed -n 1,15p > /home/cloudera/Desktop/archivo_7.txt

```
[cloudera@quickstart ~]$ hdfs dfs -cat /raw/aerolinea.csv | sed -n 1,15p >/h
ome/cloudera/Desktop/ejercicio_7.txt
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart Desktop]$ cat ejercicio_7.txt
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueC
arrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,De
pDelay,Origin,Dest,Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverte
d,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
[cloudera@quickstart Desktop]$
```

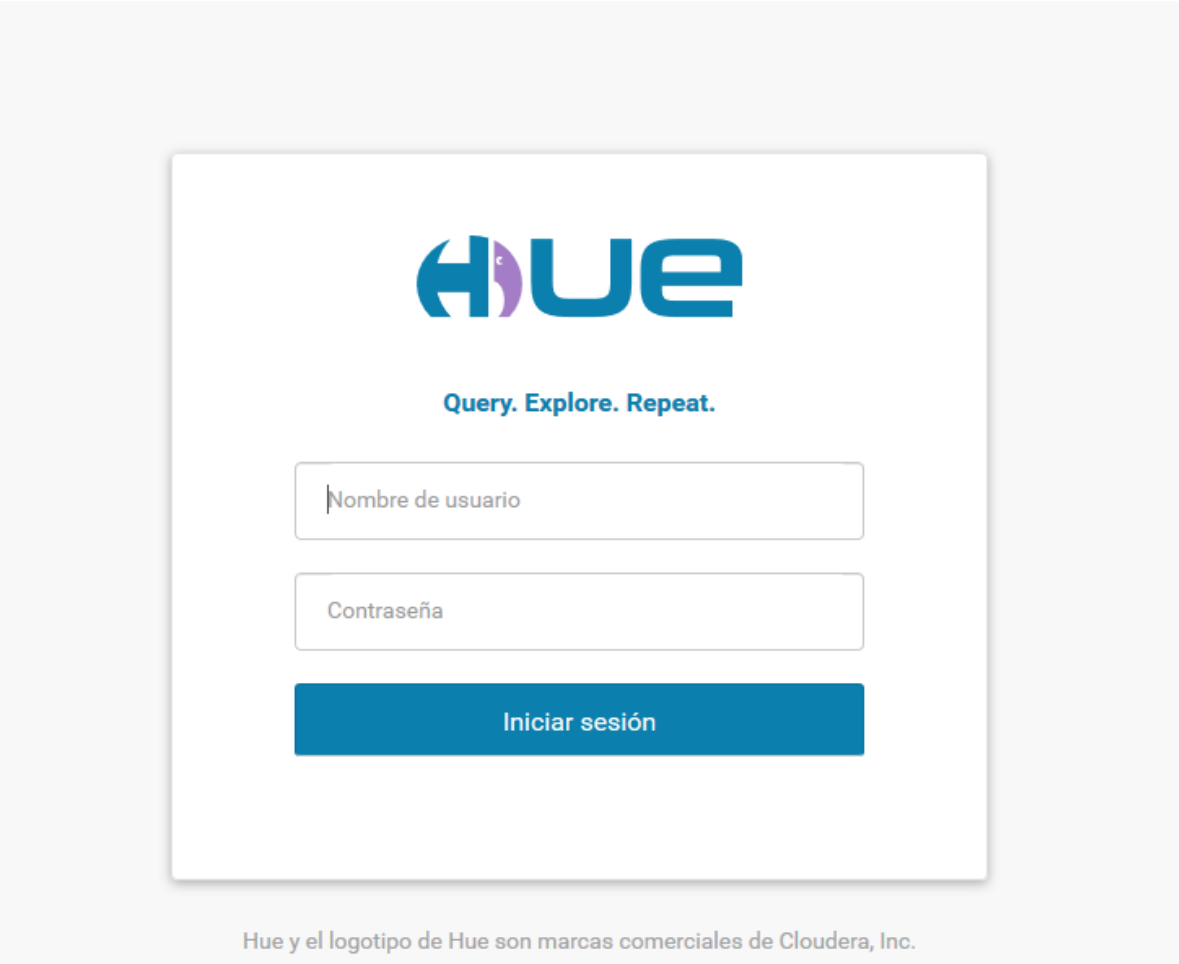
15.- Cree los directorios **master** y **stagin** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso.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

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

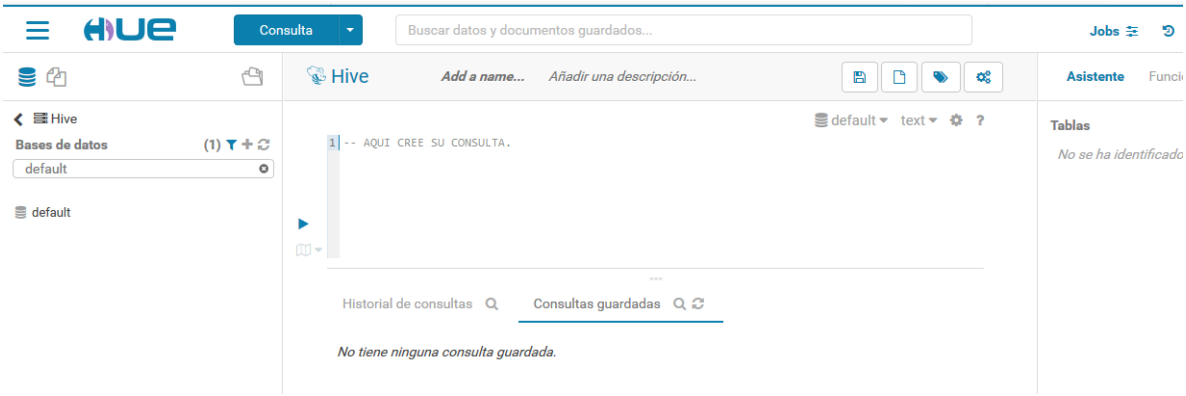
```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /master
[cloudera@quickstart ~]$ hdfs dfs -mkdir /stagin
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 9 items
drwxrwxrwx   - hdfs     supergroup          0 2017-07-19 05:34 /benchmarks
drwxr-xr-x   - hbase    supergroup          0 2018-06-22 10:30 /hbase
drwxr-xr-x   - cloudera supergroup          0 2018-06-25 10:31 /master
drwxr-xr-x   - cloudera supergroup          0 2018-06-25 10:09 /raw
drwxr-xr-x   - solr     solr                0 2017-07-19 05:37 /solr
drwxr-xr-x   - cloudera supergroup          0 2018-06-25 10:32 /stagin
drwxrwxrwt   - hdfs     supergroup          0 2018-06-22 09:12 /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
[cloudera@quickstart ~]$
```

```
[cloudera@quickstart Desktop]$ sudo -u hdfs hdfs dfs -chmod 760 /raw/aeroline
a.csv
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rwa/aerolinea.csv
ls: `/rwa/aerolinea.csv': No such file or directory
[cloudera@quickstart Desktop]$ hdfs dfs -ls /raw/aerolinea.csv
-rwxrw----   1 cloudera supergroup 12029208594 2018-06-25 10:09 /raw/aeroline
a.csv
[cloudera@quickstart Desktop]$
```

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.

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.

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.

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.

Es un bit permite evitar que un usuario pueda borrar ficheros y /o directorios de otro usuario, se podrá acceder al él y modificar, pero no se va a poder borrar nada. Se aplica al tercer grupo de permisos (Otros) y se sustituye el permiso de ejecución por una T

https://enekoamieva.com/permisos-suid-sgid-y-sticky-bit/

```
[cloudera@quickstart Desktop]$ chmod o+t archivo_5.txt
[cloudera@quickstart Desktop]$ ls -l archivo_5.txt
-rw-rw-r-T 1 cloudera cloudera 5046 Aug 25  2018 archivo_5.txt
[cloudera@quickstart Desktop]$
```

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

Las bases de datos NoSQL son bases de datos no relacionales optimizadas para modelos de datos sin esquema fotografías, audio, video, etc. También conocidas por su facilidad de desarrollo y resiliencia (capacidad de un sistema tecnológico de soportar y recuperarse ante desastres y perturbaciones) . Estos tipos de bases de datos están optimizados para aplicaciones que requieren grandes volúmenes de datos, baja latencia y modelo de datos flexibles.

Fuente: https://aws.amazon.com/es/nosql/

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

El NoHup es un comando que permite tener una ejecución de un comando a pesar de que se haya salido de terminal ya que su función principal es que ejecuta de forma independiente a la sesión. En esencia lo que hace es ignorar la señal HUP (señal que se envía al proceso cuando la terminal que lo está ejecutando se cierra.

http://rm-rf.es/nohup-mantiene-ejecucion-comando-pese-salir-terminal/

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 en los cuales tenemos que fijarnos para saber el uso de nuestra memoria RAM o cuánto está disponible es en dos campos: Free o libre (este campo nos muestra cuánto espacio es el que no estamos usando para ningún proceso o programa) y Almacenamiento o Available (ya que éste muestra el espacio libre que tenemos para poder inicializar nuevos programas)

https://geekland.eu/consumo-de-memoria-ram-en-linux/

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**?

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

Apache tiene en su distribución a Apache Hadoop y Cloudera tiene uno denominado Cloudera Hadoop que, aunque obviamente éste último viene de la misma distribución de Apache, contiene más herramientas o tecnologías como lo son Cloudera Search, Impala, Cloudera Manager y Navigator. Con éstas ventajas Cloudera termina siendo un producto patentado y de código abierto.

https://www.quora.com/What-is-the-difference-between-Apache-Hadoop-and-Cloudera-in-big-data

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

En Windows existen los archivos FAT (File Allocation Table) esta tabla se mantiene en el disco duro de la máquina y tiene un mapa de la unidad. También está el FAT16 fue la primera versión de Windows, pero se volvió obsoleto por la poca capacidad de soportar grandes volúmenes de archivos y por su capacidad de disco de 4GB. El FAT32 es una tabla de localización de 32 bits, su desventaja es que, por contener una gran cantidad de archivos, se le ve obligada a realizar fragmentaciones haciendo que la búsqueda sea más lenta. NTFS (New Tecnology File System) permite los accesos a los archivos por medio de permisos, no tiene compatibilidad en Linux ya que solo puede leerlos.

En Linux existe el ext2 era el sistema estándar en Linux, permite particiones de disco de hasta 4TB y tiene una buena estabilidad. El ext3 es la versión mejorada de la ext2 tiene una gran estabilidad y mantenimiento ya que tiene una previsión de pérdida de datos ya sea por apagones o por fallas de disco, la única desventaja es que no se pueden recuperar los archivos que se hayan borrado. Ext4 es la última versión de los ficheros ext, son más eficientes y tienen una ampliación de los tamaños de los ficheros. ReiserFS es el sistema de archivos de última generación de Linux, la organización de los ficheros agiliza las operaciones entre ellos. Swap son los ficheros de partición de Intercambio, los sistemas de Linux lo utilizan para cargar los programas sin saturar a la memoria RAM

https://prezi.com/2yhpyk6dlr_v/sistema-de-archivos-en-windows-linux-y-mac/

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

SerDe es una combinación de Serializer y Deseralizer, el primero toma un objeto en Java y lo convierte en algo que Hive puede escribir a HDFS y el segundo toma una representación binaria o string y lo convierte en un objeto de Java elcual lo puede manejar Hive, la relación entre SerDe

y Hive es que la interfaz de la primera instruye sobre la manera en la que se debe procesar un registro en Hive.

Impala no soporta el servicio SerDe.

https://unpocodejava.com/2013/01/24/apache-hive-y-serde/

https://blog.cloudera.com/blog/2012/12/how-to-use-a-serde-in-apache-hive/

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

Bigtable es un mapa distribuido ordenado y contiene tres dimensiones: filas, columnas y marca temporal. Es un sistema que divide los datos en columnas para almacenar toda la información en tablas compuestas por celdas. Tiene un alto rendimiento, es un interfaz de código abierto, el coste no es tanto como las otras alternativas NoSQL que se encuentran en el mercado.

El Bigquery es la solución que tiene Google para el Big Data, es un sitio web que permite consultar y almacenar grandes volúmenes de datos en cuestión de milisegundos

https://bbvaopen4u.com/es/actualidad/bigtable-el-servicio-de-base-de-datos-nosql-con-el-que-google-quiere-dominar-los-big-data

http://www.doctormetrics.com/2015/05/04/consultas-google-bigquery/#.WzEWP1VKjtQ

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

Data Lake es un entorno de datos compartidos en su formato original, aprovecha las herramientas y tecnologías de Big Data y comprende múltiples repositorios. El Data Lake utiliza una arquitectura plana para almacenar los datos, es decir, no tiene necesidad de guardarlos en carpetas o ficheros.

El Data Warehouse es una colección de datos la cual es variante en el tiempo, no es volátil e integrada y ayuda, gracias a su organización por el manejo de datos por temas concretos a la toma de decisiones de una empresa u organización

https://www.powerdata.es/data-lake

https://colombiadigital.net/actualidad/articulos-informativos/item/9814-que-es-un-data-warehouse-y-que-beneficios-aporta-a-las-organizaciones.html

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

CODA se desarrolló en 1987 y es para Linux.

Open gfs es un sistema de archivos en cluster.

GlusterFS es un sistema de archivos multiescalable que opera a nivel de usuario, pudiendo utilizar para la comunicación de los servidores redes TCP, tiene una estructura cliente-servidor.

https://es.slideshare.net/angyespinosa86/sistema-de-archivos-distribuidos

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: