

Tarea de Vacaciones

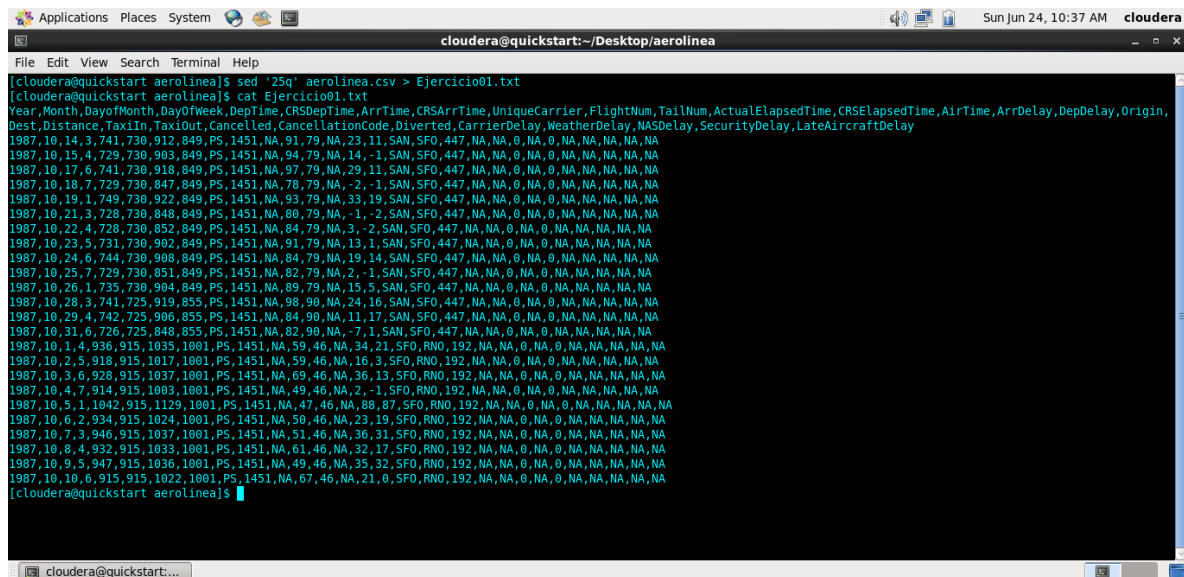
Para esta tarea se deberá responder una serie de preguntas de temas que se han abordado en el curso, el objetivo consiste entonces en reforzar los conocimientos e indagar en otros nuevos que, aunque no forman parte explícita del temario, le servirán al estudiante para incursionar en temas Big Data a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

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.

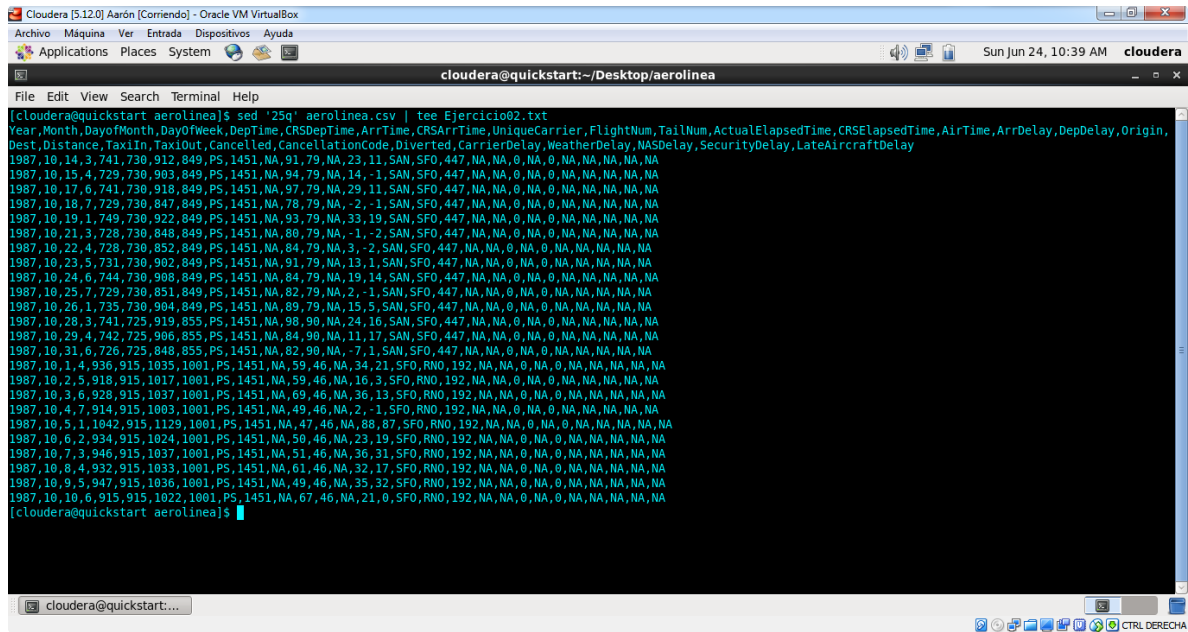
sed '25q' aerolineas.csv > Ejercicio01.txt



```
[cloudera@quickstart aerolineas]$ sed '25q' aerolineas.csv > Ejercicio01.txt
[cloudera@quickstart aerolineas]$ cat Ejercicio01.txt
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,Origin,
Dest,Distance,TaxiIn,TaxiOut,Canceled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
1987,10,14,3,741,730,912,849,PS,1451,NA,91,79,NA,23,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,15,4,729,730,903,849,PS,1451,NA,94,79,NA,14,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,17,6,741,730,918,849,PS,1451,NA,97,79,NA,29,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,18,7,729,730,847,849,PS,1451,NA,78,79,NA,-2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,19,1,749,730,922,849,PS,1451,NA,93,79,NA,33,19,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,21,3,728,730,848,849,PS,1451,NA,80,79,NA,-1,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,24,6,744,730,908,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,28,3,741,725,919,855,PS,1451,NA,98,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,5,1,1042,915,1129,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,6,2,934,915,1024,1001,PS,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
1987,10,7,3,946,915,1027,1001,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 aerolineas]$
```

2.-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.

sed '25q' aerolineas.csv | tee Ejercicio02.txt



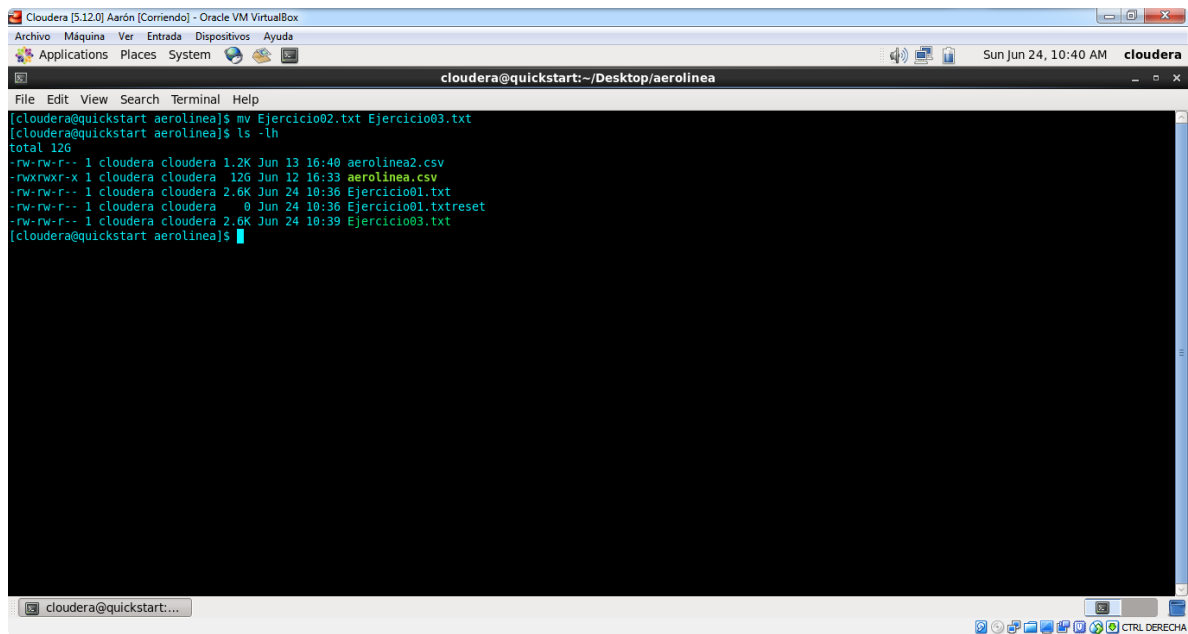
```
Cloudera [5.12.0] Aarón [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Applications Places System
Sun Jun 24, 10:39 AM cloudera

cloudera@quickstart:~/Desktop/aerolinea

[cloudera@quickstart aerolinea]$ sed '25q' aerolinea.csv | tee Ejercicio02.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
1987,10,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,24,6,744,730,988,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,28,3,741,725,919,855,PS,1451,NA,98,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,5,1,1042,915,1129,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,6,2,934,915,1024,1001,PS,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,7,3,946,915,1037,1001,PS,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,8,4,932,915,1033,1001,PS,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,9,5,947,915,1036,1001,PS,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
1987,10,10,6,915,915,1022,1001,PS,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA
[cloudera@quickstart aerolinea]$
```

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

mv Ejercicio02.txt Ejercicio03.txt



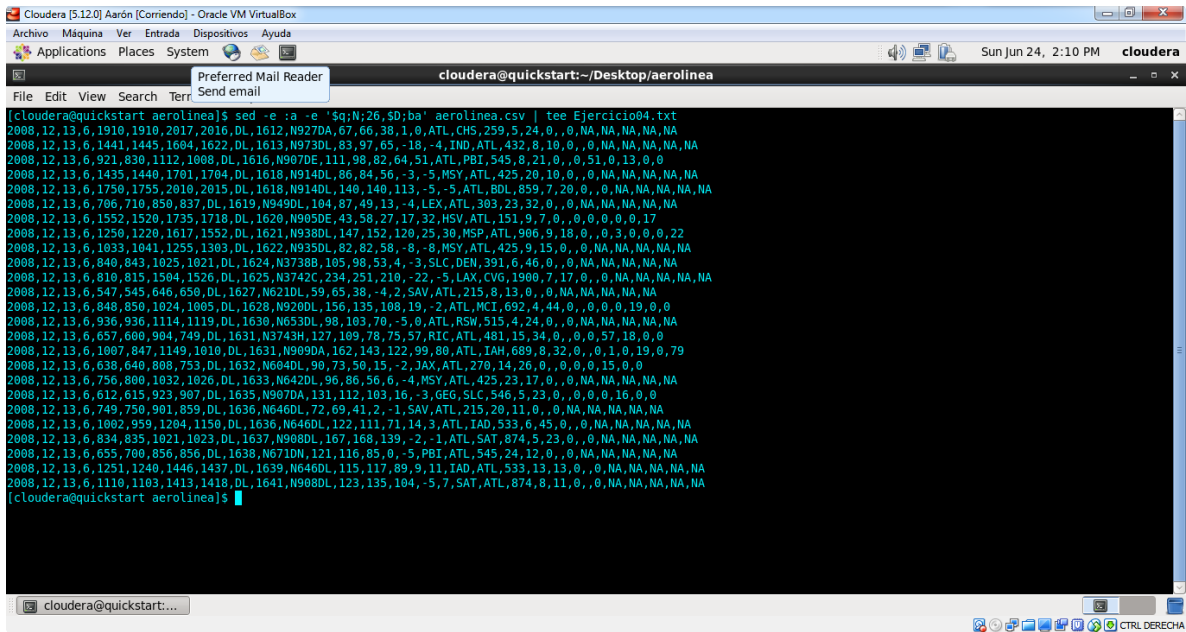
```
Cloudera [5.12.0] Aarón [Corriendo] - Oracle VM VirtualBox
Archivo Máquina Ver Entrada Dispositivos Ayuda
Applications Places System
Sun Jun 24, 10:40 AM cloudera

cloudera@quickstart:~/Desktop/aerolinea

[cloudera@quickstart aerolinea]$ mv Ejercicio02.txt Ejercicio03.txt
[cloudera@quickstart aerolinea]$ ls -lh
total 120
-rw-rw-r-- 1 cloudera cloudera 1.2K Jun 13 16:40 aerolinea2.csv
-rwxrwxr-x 1 cloudera cloudera 126 Jun 12 16:33 aerolinea.csv
-rw-rw-r-- 1 cloudera cloudera 2.6K Jun 24 10:36 Ejercicio01.txt
-rw-rw-r-- 1 cloudera cloudera 0 Jun 24 10:36 Ejercicio01.txtreset
-rw-rw-r-- 1 cloudera cloudera 2.0K Jun 24 10:39 Ejercicio03.txt
[cloudera@quickstart aerolinea]$
```

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**

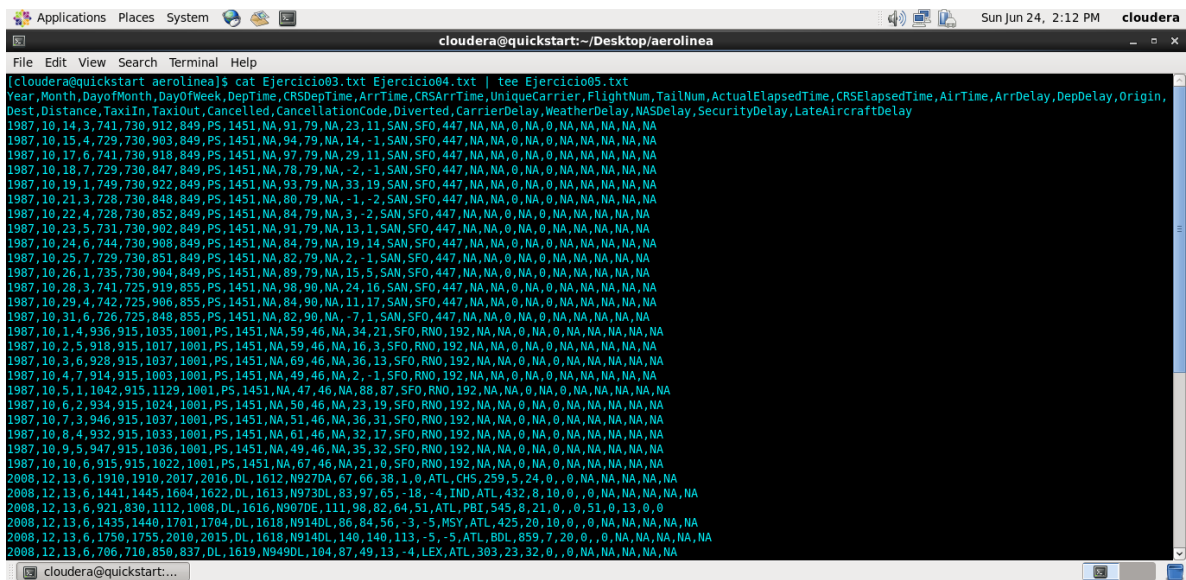
sed -e :a -e '\$q;N;26,\$D;ba' aerolineas.csv | tee Ejercicio04.txt



```
cloudera@quickstart:~/Desktop/aerolinea
[cloudera@quickstart aerolinea]$ sed -e 's;q;N;26,50;ba' aerolinea.csv | tee Ejercicio04.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,BOL,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
2008,12,13,6,1552,1550,1735,1718,DL,1620,N905DE,43,58,27,17,32,HSV,ATL,151,9,7,0,,0,0,0,0,17
2008,12,13,6,1250,1250,1617,1552,DL,1621,N938DL,147,152,120,25,30,MSR,ATL,900,9,18,0,,0,3,0,0,22
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,N37388,105,98,53,4,-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
2008,12,13,6,638,640,888,753,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
2008,12,13,6,749,750,901,850,DL,1636,N646DL,72,69,41,2,-1,SAV,ATL,215,20,11,0,,0,NA,NA,NA,NA,NA
2008,12,13,6,1002,950,1204,1150,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 aerolinea]$
```

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**

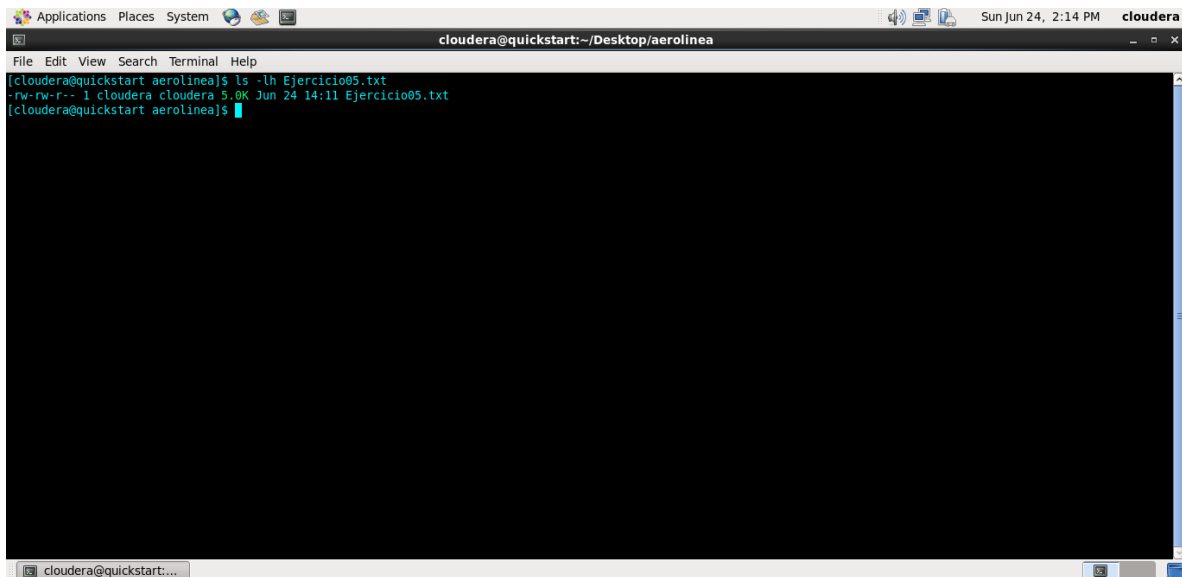
cat Ejercicio03.txt Ejercicio04.txt | tee Ejercicio05.txt



```
cloudera@quickstart:~/Desktop/aerolinea
[cloudera@quickstart aerolinea]$ cat Ejercicio03.txt Ejercicio04.txt | tee Ejercicio05.txt
Year,Month,DayOfMonth,DayOfWeek,DepTime,CRSDepTime,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
1987,10,4,7,914,915,1009,1001,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,BOL,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
```

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.

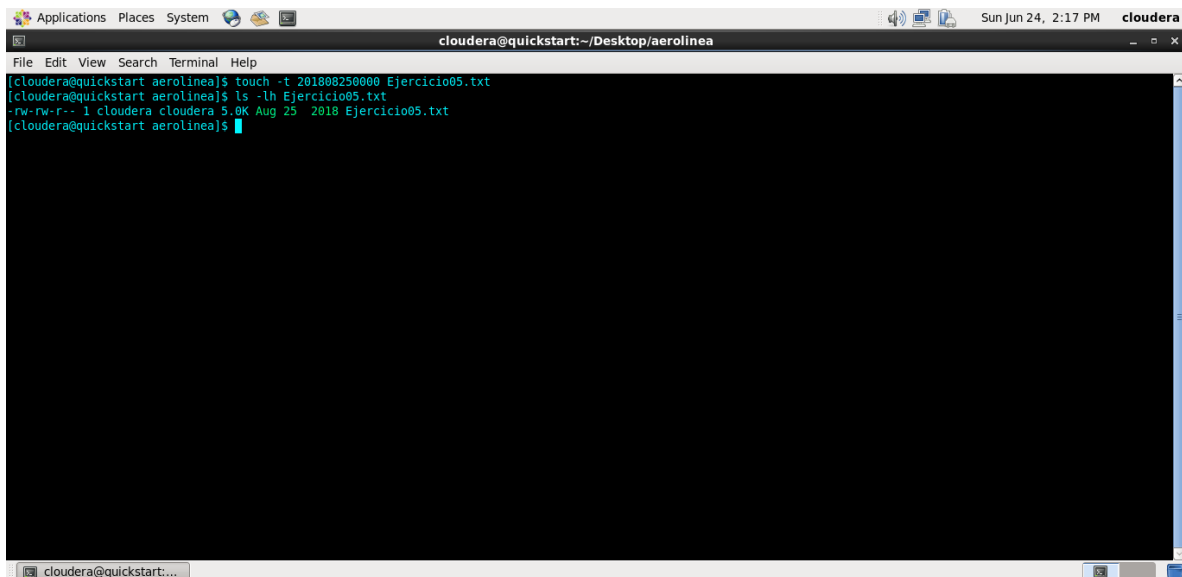
ls -lh Ejercicio05.txt



```
Applications Places System cloudera@quickstart:~/Desktop/aerolinea
File Edit View Search Terminal Help
[cloudera@quickstart aerolinea]$ ls -lh Ejercicio05.txt
-rw-rw-r-- 1 cloudera cloudera 5.0K Jun 24 14:11 Ejercicio05.txt
[cloudera@quickstart aerolinea]$
```

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 -t 201808250000 Ejercicio05.txt



```
Applications Places System cloudera@quickstart:~/Desktop/aerolinea
File Edit View Search Terminal Help
[cloudera@quickstart aerolinea]$ touch -t 201808250000 Ejercicio05.txt
[cloudera@quickstart aerolinea]$ ls -lh Ejercicio05.txt
-rw-rw-r-- 1 cloudera cloudera 5.0K Aug 25 2018 Ejercicio05.txt
[cloudera@quickstart aerolinea]$
```

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

```
Applications Places System cloudera@quickstart:~/Desktop/aerolinea
File Edit View Search Terminal Help
[cloudera@quickstart aerolinea]$ 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        : 2608.920
cache size     : 4096 KB
physical id    : 0
siblings       : 2
core id        : 0
cpu cores      : 2
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 ht syscall nx rdtscp lm constant_tsc rep_good xtopol
oggy nonstop_tsc unfair_spinlock pni pclmulqdq ssse3 cx16 pcid sse4_1 sse4_2 movbe popcnt aes xsave avx rdrand hypervisor lahf_lm abm 3dnowprefetch fsgsbase avx2 invpcid
rdseed
bogomips       : 5217.84
clflush size   : 64
cache alignment : 64
address sizes   : 39 bits physical, 48 bits virtual
power management:

processor       : 1
vendor_id      : GenuineIntel
cpu family     : 6
model          : 78
model name     : Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz
```

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:

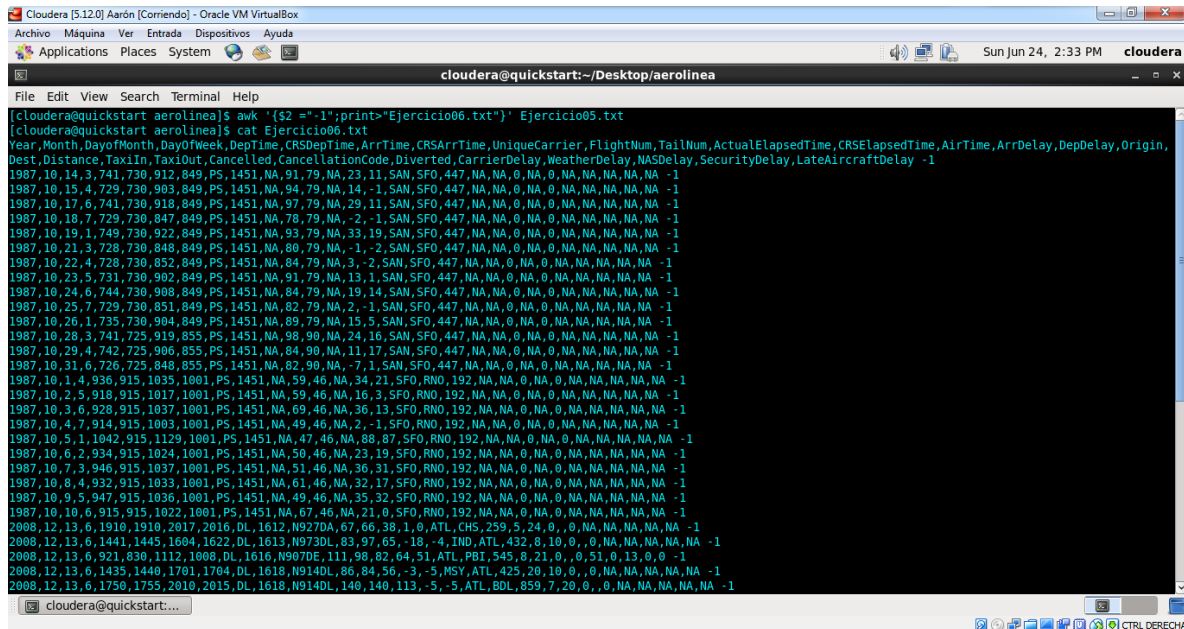
```
aaron@aaron-trabajo-vb:~/Descargas$ cat muestra.txt
a,c,b,d,e,f
g,h,i,j,k,l
m,n,o,p,q,r
aaron@aaron-trabajo-vb:~/Descargas$ awk
be
ik
oq
aaron@aaron-trabajo-vb:~/Descargas$
```

awk -F, '{OFS=",";print \$3,\$5}' Ejercicio05.txt

```
Applications Places System cloudera@quickstart:~/Desktop/aerolinea
File Edit View Search Terminal Help
[cloudera@quickstart aerolinea]$ awk -F, '{OFS=",";print $3,$5}' Ejercicio05.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
13,706
13,1552
```

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 '{ \$2 = "-1"; print>" Ejercicio06.txt" }' Ejercicio05.txt



```
Cloudera [512.0] Aaron [Corriendo] - Oracle VM VirtualBox
Archivo Maquina Ver Entrada Dispositivos Ayuda
Applications Places System
cloudera@quickstart:~/Desktop/aerolinea

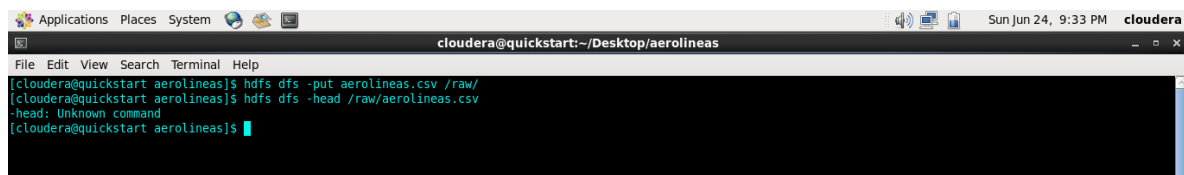
File Edit View Search Terminal Help
[cloudera@quickstart aerolinea]$ awk '{ $2 = "-1"; print>" Ejercicio06.txt" }' Ejercicio05.txt
[cloudera@quickstart aerolinea]$ cat Ejercicio06.txt
Year,Month,DayOfMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,Origin,
Dest,Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay -1
1987,10,14,3,741,730,912,849,PS,1451,NA,91,79,NA,22,11,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA -1
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 -1
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 -1
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 -1
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 -1
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 -1
1987,10,22,4,728,730,852,849,PS,1451,NA,84,79,NA,3,-2,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,23,5,731,730,902,849,PS,1451,NA,91,79,NA,13,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,24,6,744,730,908,849,PS,1451,NA,84,79,NA,19,14,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,25,7,729,730,851,849,PS,1451,NA,82,79,NA,2,-1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,26,1,735,730,904,849,PS,1451,NA,89,79,NA,15,5,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,28,3,741,725,919,855,PS,1451,NA,90,90,NA,24,16,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,29,4,742,725,906,855,PS,1451,NA,84,90,NA,11,17,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,31,6,726,725,848,855,PS,1451,NA,82,90,NA,-7,1,SAN,SFO,447,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,1,4,936,915,1035,1001,PS,1451,NA,59,46,NA,34,21,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,2,5,918,915,1017,1001,PS,1451,NA,59,46,NA,16,3,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,3,6,928,915,1037,1001,PS,1451,NA,69,46,NA,36,13,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,4,7,914,915,1003,1001,PS,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,5,1,1042,915,1129,1001,PS,1451,NA,47,46,NA,88,87,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,6,2,934,915,1024,1001,PS,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,7,3,946,915,1037,1001,PS,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,8,4,932,915,1033,1001,PS,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,9,5,947,915,1036,1001,PS,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
1987,10,10,6,915,915,1022,1001,PS,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA,NA -1
2008,12,13,6,1910,1910,2017,2016,DL,1612,N927DA,67,66,30,1,0,ATL,CHS,250,5,24,0,0,NA,NA,NA,NA,NA,NA -1
2008,12,13,6,1441,1445,1604,1622,DL,1613,N973DL,83,97,65,-13,-4,IND,ATL,432,8,10,0,0,NA,NA,NA,NA,NA,NA -1
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 -1
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 -1
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 -1
```

SECCION 2. HDFS Y HIVE

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

hdfs dfs -head /raw/aerolineas.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.



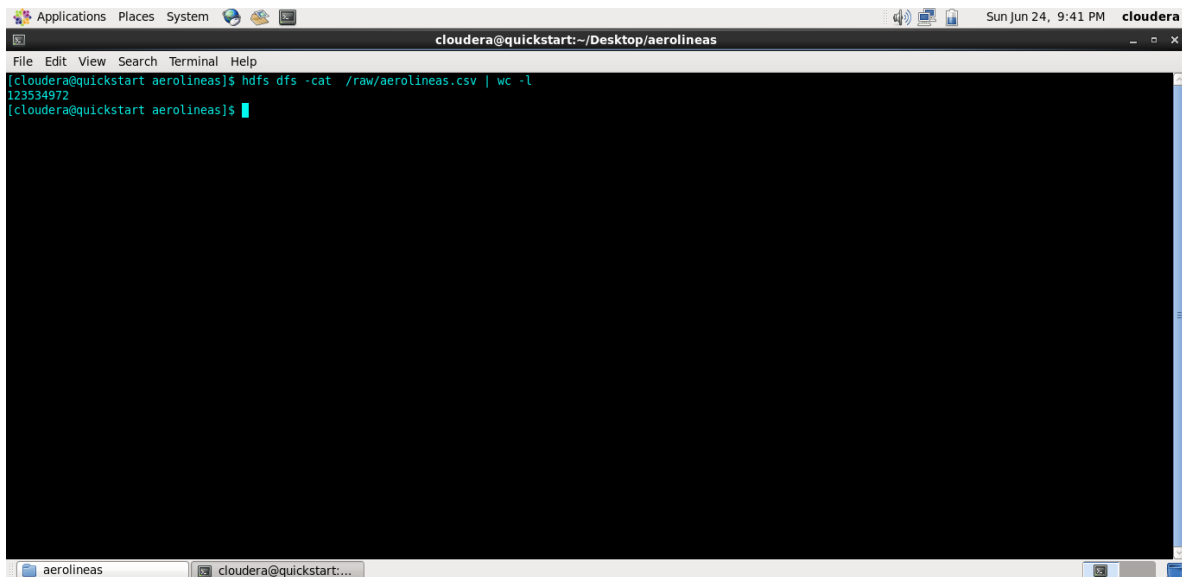
```
Applications Places System
cloudera@quickstart:~/Desktop/aerolineas

File Edit View Search Terminal Help
[cloudera@quickstart aerolineas]$ hdfs dfs -put aerolineas.csv /raw/
[cloudera@quickstart aerolineas]$ hdfs dfs -head /raw/aerolineas.csv
-head: Unknown command
[cloudera@quickstart aerolineas]$
```

El comando head no pertenece a los comandos del HDFS, si bien es funcional en Linux, dentro del HDFS no existe.

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

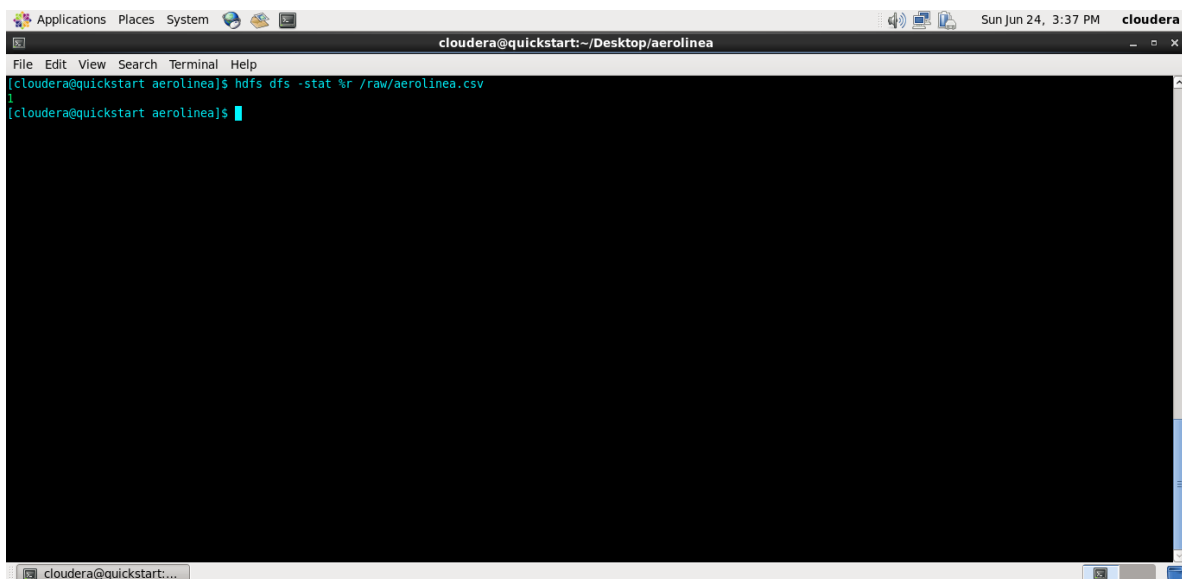
hdfs dfs -cat /raw/aerolineas.csv | wc -l



```
Applications Places System cloudera@quickstart:~/Desktop/aerolineas
File Edit View Search Terminal Help
[cloudera@quickstart aerolineas]$ hdfs dfs -cat /raw/aerolineas.csv | wc -l
123534972
[cloudera@quickstart aerolineas]$
```

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/aerolineas.csv



```
Applications Places System cloudera@quickstart:~/Desktop/aerolineas
File Edit View Search Terminal Help
[cloudera@quickstart aerolineas]$ hdfs dfs -stat %r /raw/aerolineas.csv
1
[cloudera@quickstart aerolineas]$
```

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/aerolineas.csv | sed '15q' >
/home/cloudera/Desktop/aerolineas/Ejercicio07.txt

```
Applications Places System cloudera@quickstart:~/Python-3.6.3
File Edit View Search Terminal Help
[cloudera@quickstart Python-3.6.3]$ hdfs dfs -cat /raw/aerolineas.csv | sed '15q' > /home/cloudera/Desktop/aerolineas/Ejercicio07.txt
cat: Unable to write to output stream.
[cloudera@quickstart Python-3.6.3]$ cat /home/cloudera/Desktop/aerolineas/Ejercicio07.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,90,90,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 Python-3.6.3]$
```

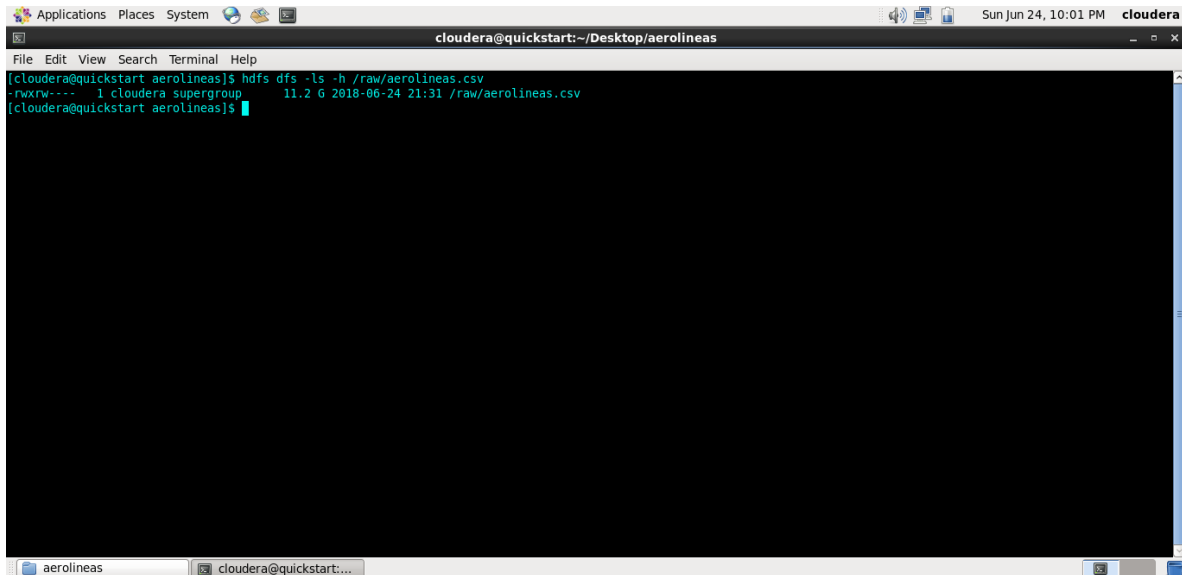
15.- Cree los directorios **master** y **stagin** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso.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

hdfs dfs -mkdir /master

hdfs dfs -mkdir /stagin

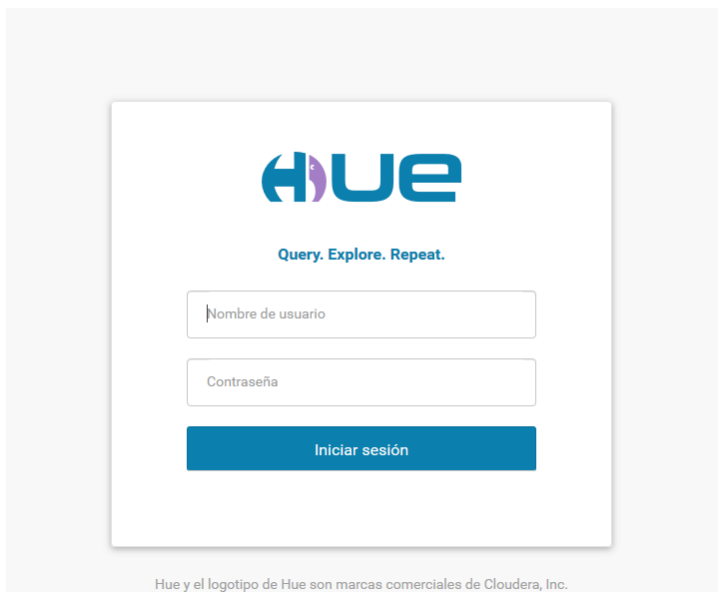
```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -ls h /
ls: 'h': No such file or directory
Found 9 items
drwxrwxrwx - hdfs supergroup 0 2017-07-19 05:34 /benchmarks
drwxr-xr-x - hbase supergroup 0 2018-06-24 17:25 /hbase
drwxr-xr-x - hdfs supergroup 0 2018-06-13 16:27 /master
drwxr-xr-x - hdfs supergroup 0 2018-06-13 16:51 /raw
drwxr-xr-x - solr solr 0 2017-07-19 05:37 /solr
drwxr-xr-x - hdfs supergroup 0 2018-06-13 16:27 /stagin
drwxrwxrwt - hdfs supergroup 0 2018-06-24 13:55 /tmp
drwxr-xr-x - hdfs supergroup 0 2018-06-13 17:02 /user
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /var
[cloudera@quickstart ~]$
```

hdfs dfs -chmod 760 /raw/aerolineas.csv

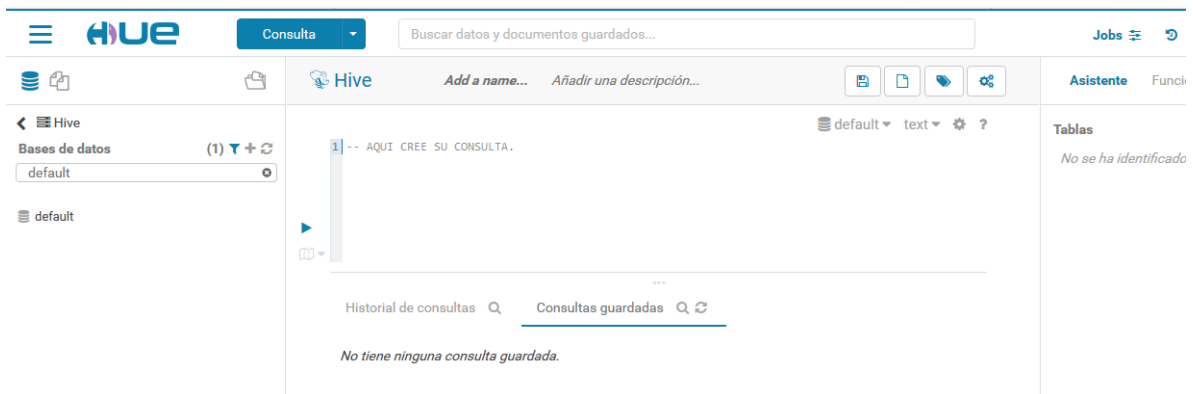


```
cloudera@quickstart:~/Desktop/aerolineas
File Edit View Search Terminal Help
[cloudera@quickstart aerolineas]$ hdfs dfs -ls -h /raw/aerolineas.csv
-rwxrw---- 1 cloudera supergroup 11.2 G 2018-06-24 21:31 /raw/aerolineas.csv
[cloudera@quickstart aerolineas]$
```

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,
```

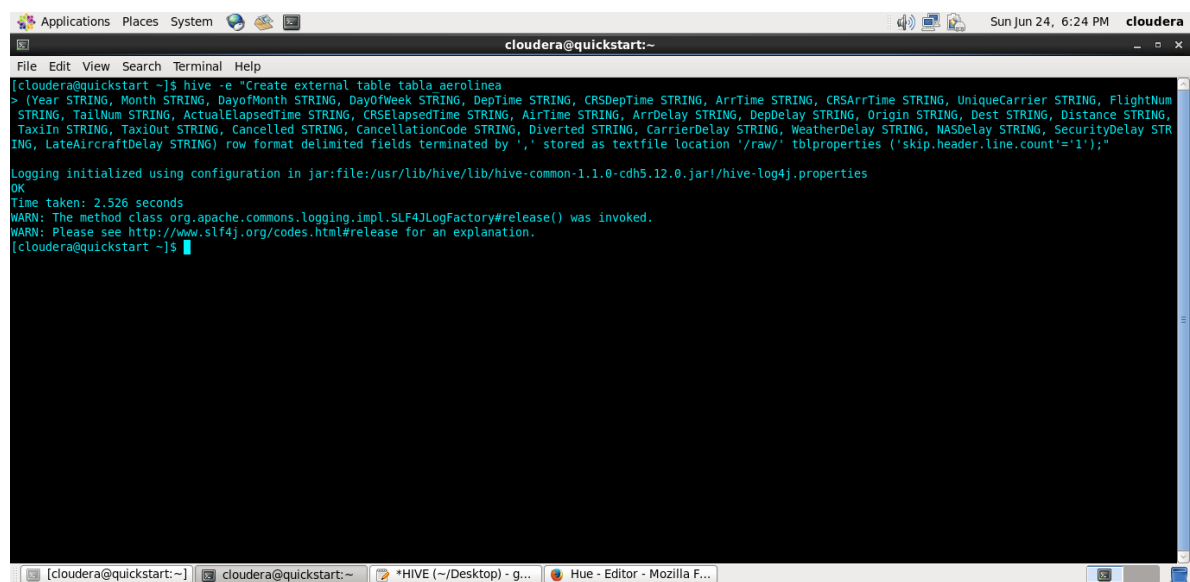
SecurityDelay STRING,
LateAircraftDelay STRING)

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

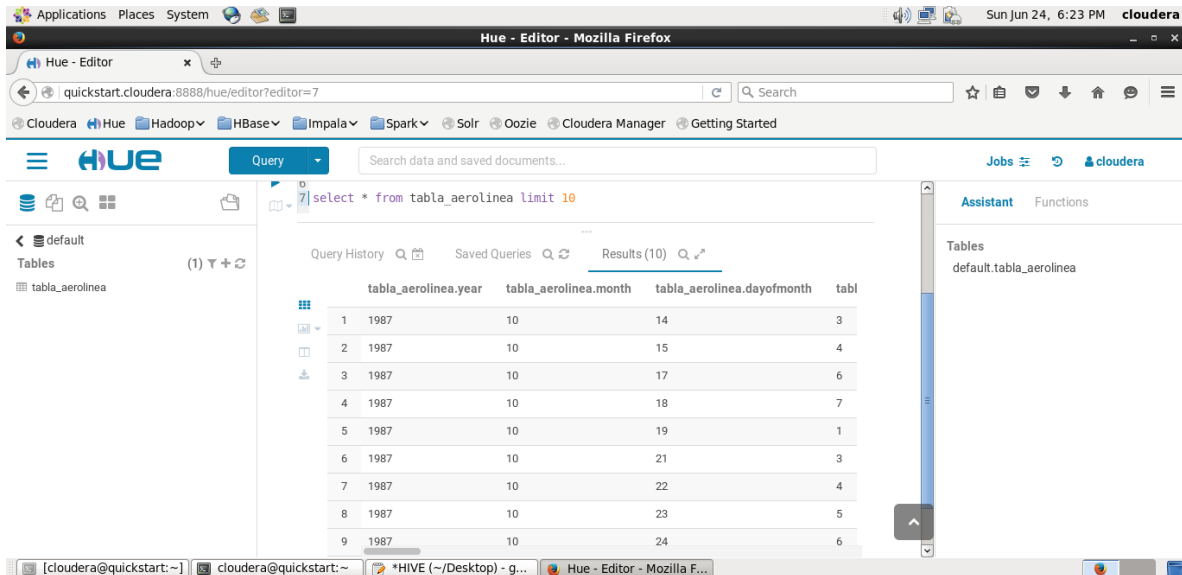
En el código anterior **NO** existe una forma de omitir los encabezados por lo que es su deber encontrar esa manera, incluirla en el código y crear la tabla.

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.

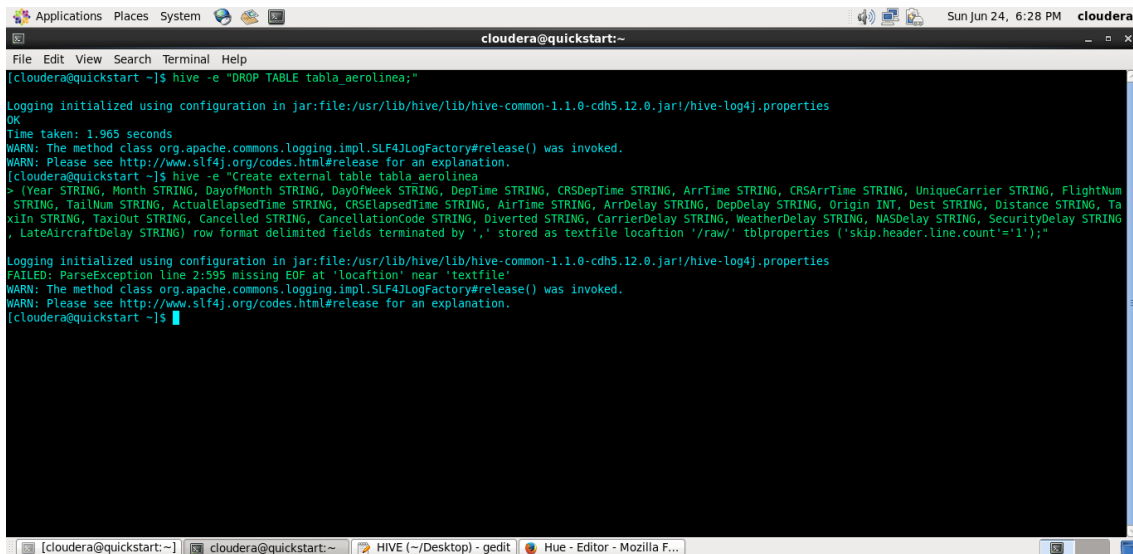
```
hive -e "CREATE EXTERNAL TABLE tabla_aerolinea (Year STRING,Month STRING,DayofMonth  
STRING,DayOfWeek STRING,DepTime STRING,CRSDepTime STRING,ArrTime STRING,CRSArrTime  
STRING,UniqueCarrier STRING,FlightNum STRING,TailNum STRING,ActualElapsedTime  
STRING,CRSElapsedTime STRING,AirTime STRING,ArrDelay 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');"
```



```
Applications Places System cloudera@quickstart:~
cloudera@quickstart:~$ hive -e "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 STR
ING, LateAircraftDelay STRING) row format delimited fields terminated by ',' stored as textfile 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: 2.526 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:~$
```



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.



No se consigue la creación de la tabla porque no coinciden los datos de la columna con el tipo de dato INT.

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.

hive -e "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

The image shows a terminal window titled 'cloudera@quickstart:~'. The terminal displays the following commands and output:

```
cloudera@quickstart ~]$ hive -e "create external table tabla aerolineas ( Year STRING, Month STRING, DayOfMonth STRING, DayOfWeek STRING, DepTime STRING, CRSDepTime STRING, ArrTime STRING, CRSArrTime STRING, UniqueCarrier STRING, FlightNum STRING, TailNum STRING, ActualElapsedTime STRING, CRSElapsedTime STRING, AirTime STRING, ArrDelay STRING, DepDelay STRING, Origin STRING, Dest STRING, Distance STRING, TaxiIn STRING, TaxiOut STRING, Cancelled STRING, CancellationCode STRING, Diverted STRING, CarrierDelay STRING, WeatherDelay STRING, NASDelay STRING, SecurityDelay STRING, LateAircraftDelay STRING, Adicional STRING ) row format delimited fields terminated by ',' stored as textfile location '/raw/' tblproperties ('skip.header.line.count'='1');"
```

The output of the command is:

```
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: 1.646 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 ~]$
```

The terminal window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The status bar at the bottom shows the terminal is running on 'cloudera@quickstart:~' and the window title is 'HIVE (~/.Desktop) - gedit'.

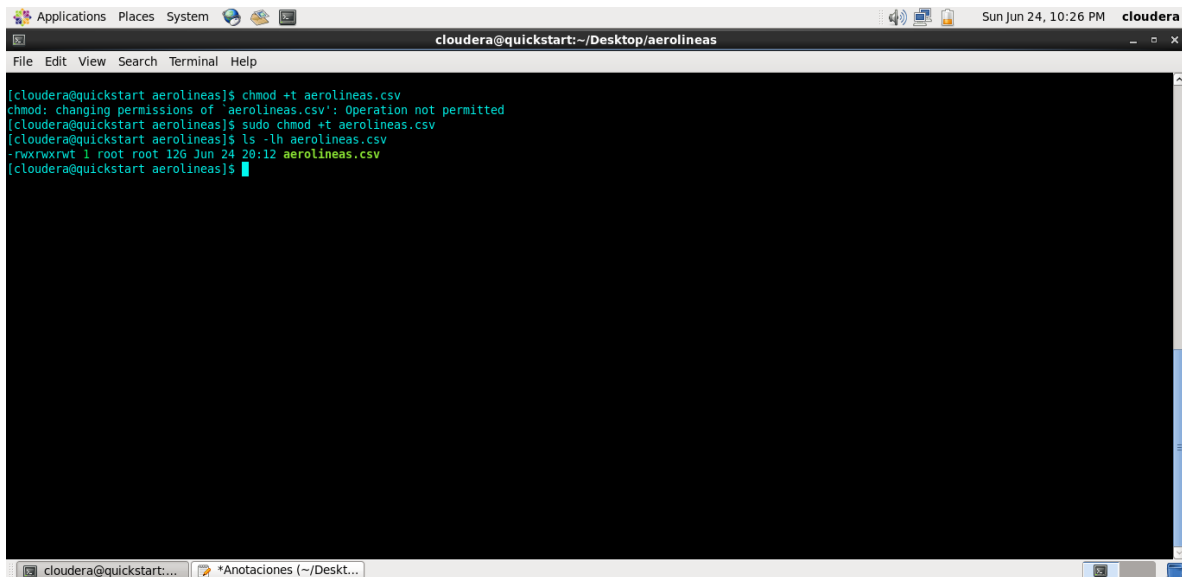
19.- En esta tabla anterior inserte un renglón a la tabla con todos los valores iguales a “NA” (tiene que investigar cómo añadir elementos a la tabla), y luego después de la inserción del elemento indague en qué parte del HDFS se ha guardado ese nuevo elemento.

SECCIÓN 3. PREGUNTAS ABIERTAS

La funcionalidad del sticky bit depende del tipo de archivo, para directorios evita que los usuarios sin privilegios pueda cambiar el nombre o borrar el contenido de un archivo en el directorio a menos que sea el usuario o root. Para los archivos comunes guarda la imagen del texto por lo que puede cargar más rápidamente. Para poder modificarlo se utiliza alguno de los siguientes comandos:

- `chmod o+t` or `chmod +t`
- `chmod 1757` donde 1 significa sticky bit.

<https://www.linuxnix.com/sticky-bit-set-linux/>



```

[cloudera@quickstart aerolineas]$ chmod +t aerolineas.csv
chmod: changing permissions of 'aerolineas.csv': Operation not permitted
[cloudera@quickstart aerolineas]$ sudo chmod +t aerolineas.csv
[cloudera@quickstart aerolineas]$ ls -lh aerolineas.csv
-rwxrwxrwt 1 root root 126 Jun 24 20:12 aerolineas.csv
[cloudera@quickstart aerolineas]$

```

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

Las tecnologías NoSQL por su complejidad tienden a dividirse en cuatro grupos:

1. **Bases de datos Clave-Valor:** Son modelos de datos sencillos que tienen una clave indexada a un valor. Tienen un tiempo de respuesta rápido y disponibilidad total.
2. **Orientadas a Documentos:** Utilizan el modelo de documento, normalmente en formato JSON para almacenar y consultar información. Gestiona información con estructuras jerárquicas complejas. Dada su flexibilidad del esquema de datos las convierten en más versátiles y de propósito general.
3. **Orientadas a grafos:** El modelo de datos se centra en entidades (nodos del grado) y las relaciones entre estas (aristas). Se tienen que recorrer las uniones entre estas relaciones, esto permite hacerlo con gran velocidad
4. **Orientadas a columnas:** Son similares a bases de datos relacionales, pero un registro puede contener cualquier número de columnas o familia de columnas

<https://www.paradigmadigital.com/techbiz/breve-introduccion-las-tecnologias-nosql/>

Hive se utiliza para realizar tareas intensivas de datos. Es una infraestructura de datawarehouse construida sobre la plataforma de Hadoop. Permite grandes conjuntos de datos almacenados en HDFS y Hadoop. Ofrece un lenguaje similar al SQL y lee y convierte las consultas a MapReduce, Apache Tez y Spark. Soporta varios tipos de almacenamientos como texto plano, RCFile, HBase, ORC.

Impala es un motor de SQL de procesamiento masivo paralelo (MPP), se puede integrar en el sistema de HADOOP. Soporta archivos como texto plano: LZO, Avro, RCFile, Parquet,

<https://data-flair.training/blogs/impala-vs-hive/>

Consideron que ambos pueden soportar los diferentes tipos de grupos 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?

Una señal HUP es la señal que se envía a un proceso cuando la terminal que lo controla se cierra. Entonces la señal NOHUP es que el proceso se siga ejecutando aunque la terminal se haya cerrado ya que el proceso se ejecuta de forma independiente a la sesión. Se puede redirigir el error a un Shell determinado.

<https://linuxide.com/how-tos/example-how-to-use-linux-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.

Línea 1: Muestra la información cómo la memoria física está siendo utilizada.

Línea 2: Muestra cómo se está utilizando la memoria con caché.

Línea 3: Muestra el uso de la memoria swap la memoria swap cómo está siendo utilizada y entre

<https://www.networkworld.com/article/2722141/it-management/making-sense-of-memory-usage-on-linux.html>

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

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

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

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

chmod 575 objetivo.txt → Permisos de lectura y ejecución para el usuario, todos los permisos para el grupo y permisos de ejecución y lectura a otros.

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

Cloudera es una plataforma escalables, flexible, e integrada que facilita el manejo de volúmenes de datos de crecimientos rápido y variedades de datos. Cloudera implementa y administra Apache Hadoop para los proyectos relacionados, manipula y analiza los datos y mantiene los datos seguros y protegidos.

<https://www.cloudera.com/documentation/enterprise/5-2-x/PDF/cloudera-introduction.pdf>

Hadoop es un ecosistema de componentes de open source que cambia la manera de guardar, procesar y analizar los datos. Hadoop permite múltiples tipos carga para que los datos se ejecuten al mismo tiempo y escala masiva.

<https://www.cloudera.com/products/open-source/apache-hadoop.html>

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

Para GNU/Linux existen 7 diferentes tipos de archivos:

- 1. Archivo Regular:** Estos son archivos como; texto, Imágenes, archivos binarios, librerías compartidas, csv, etc.
- 2. Directorios:** Son archivos donde guardan diferentes tipos de archivos.
- 3. Dispositivos de Caracter:** Son archivos que permiten al usuario comunicarse con dispositivos exteriores (hardware)
- 4. Dispositivos de Bloque:** Al igual que los dispositivos de caracter permiten comunicarse con dispositivos exteriores usualmente discos duros, memorias, etc.
- 5. Socket de dominios locales:** Usados para la comunicación entre procesos
- 6. Pipes Named:** Usado para la comunicación de dos procesos locales.
- 7. Links Sombólicos:** Existen dos, suave y duro. Son links que el usuario asigna a un archivo.

<https://linuxconfig.org/identifying-file-types-in-linux>

Para Windows existen 2 tipos de archivos:

- 1. Públicos:** Los definen los organismos de estándares y/o son promovidos por sus organizaciones definitorias como formatos de intercambio, pueden intercambiarse entre usuarios y computadoras, son compatibles en muchas plataformas.
- 2. Privadas.** Son de un formato implementado y entendido por una sola aplicación o proveedor

[https://msdn.microsoft.com/en-us/library/windows/desktop/cc144148\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/windows/desktop/cc144148(v=vs.85).aspx)

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

SerDe es un framework para serializar y deserializar las estructuras de datos de manera eficiente y genérica. El ecosistema de Serde es cómo las estructuras de datos se serializan y deserializan junto con los formatos de datos que pueden hacer lo mismo como por ejemplo (JSON, Bincode, CBOR, YAML, etc). Serde proporciona la capa mediante estos dos grupos interactúan entre sí, permitiendo soportar cualquier estructura de datos pudiéndose ser serializada y deserializada utilizando cualquier formato de datos compatible.

<https://serde.rs/>

RELACIÓN ENTRE SERDE Y HIVE

La interfaz de deserialización toma una cadena o una representación binaria de un registro y lo traduce a objetos de Java donde Hive lo puede manipular. La serialización tomará un objeto de Java donde Hive ha estado trabajando y lo cambia a algo donde Hive lo puede escribir en HDFS .

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

RELACIÓN ENTRE SERDE E IMPALA

El cliente SerDe no es compatible con Impala.

<https://community.cloudera.com/t5/Interactive-Short-cycle-SQL/Impala-support-for-custom-serde/m-p/4185>

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

Big Table es el manejador de base de datos de Big Data NoSQL de Google, se integra con las herramientas de Open Source

<https://cloud.google.com/bigtable/>

Big Query es un almacén de datos empresariales de google de bajo coste, de gran escalabilidad y sin servidor diseñado para el análisis de los datos.

<https://cloud.google.com/bigquery/>

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

Data Lake almacena datos relacionales y datos no relacionarl (SQL y noSQL) de aplicaciones móviles, IoT (internet of things) y redes sociales. La estructura de los datos no se captura cuando se almacenan los datos. Se utilizan diferentes tipos de análisis para este tipo de almacenamiento.

Data Warehouse es una base de datos optimizada relacional para analizar datos de sistemas transaccionales. Se definen la estructura de datos para poderlos consultar más rápido.

<https://aws.amazon.com/es/big-data/datalakes-and-analytics/what-is-a-data-lake/>

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

- **MooseFS:** es un DFS desarrollado por Gemius SA
- **iRODS:** es un DFS desarrollado por el grupo de Data Intensive Cyber Enviroments (DICE)
- **Ceph:** Desarrollado por Sage Weil
- **GlusterFS:** Desarrollado por el equipo de cluster core
- **Lustre:** Sólo es para Linux bajo la licencia de GPL

<https://hal.inria.fr/hal-00789086/document>

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:

