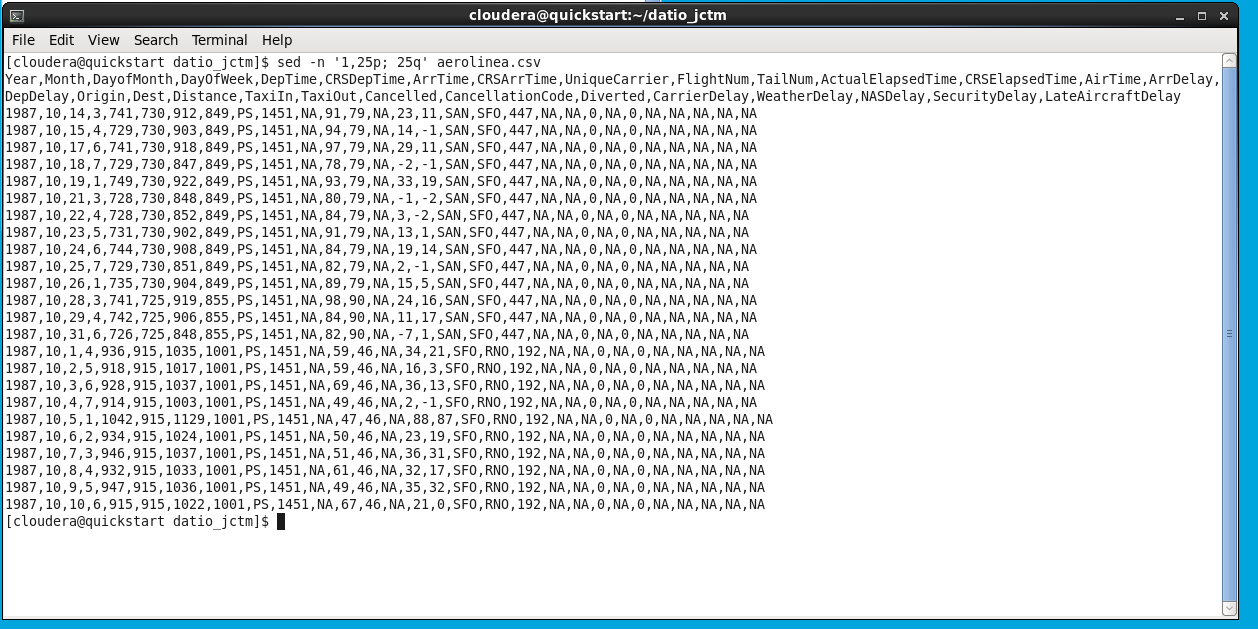
**SECCION 1. GNU/LINUX**

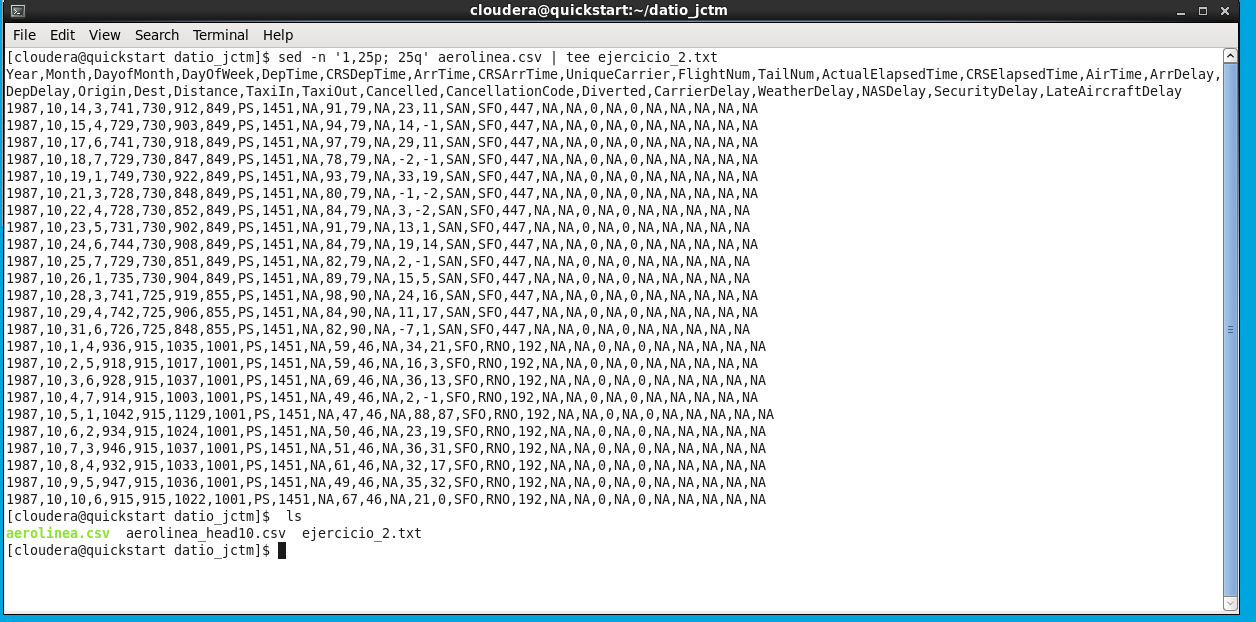
1.- Del archivo **aerolineas.csv** (el archivo descomprimido que todavía debería estar en su local y no el del HDFS) use comandos de GNU/Linux para obtener las 25 primeras líneas (incluyendo encabezado) **SIN** usar el comando head.  
Respuesta : sed -n '1,25p; 25q' aerolinea.csv



2.-Como ya se ha visto, utilizar el redireccionamiento destructivo (>) implica almacenar típicamente algún contenido en un archivo **(ej. echo "contenido" > archivo)**.   
Pero lo cierto es que con este comando no se apreciará en pantalla lo que se desea almacenar en dicho archivo, por ello es que se necesita que, con base en el comando resultado del ejercicio 1 y con la investigación del comando **tee**, por un lado el contenido se introduzca en el archivo **ejercicio\_2.txt** y por el otro se muestre en pantalla la operación.  
Caber mencionar que todo se debe registrar como una sola instrucción, es decir, no se puede ejecutar el resultado por partes, para ello tal vez quiera leer esta liga:

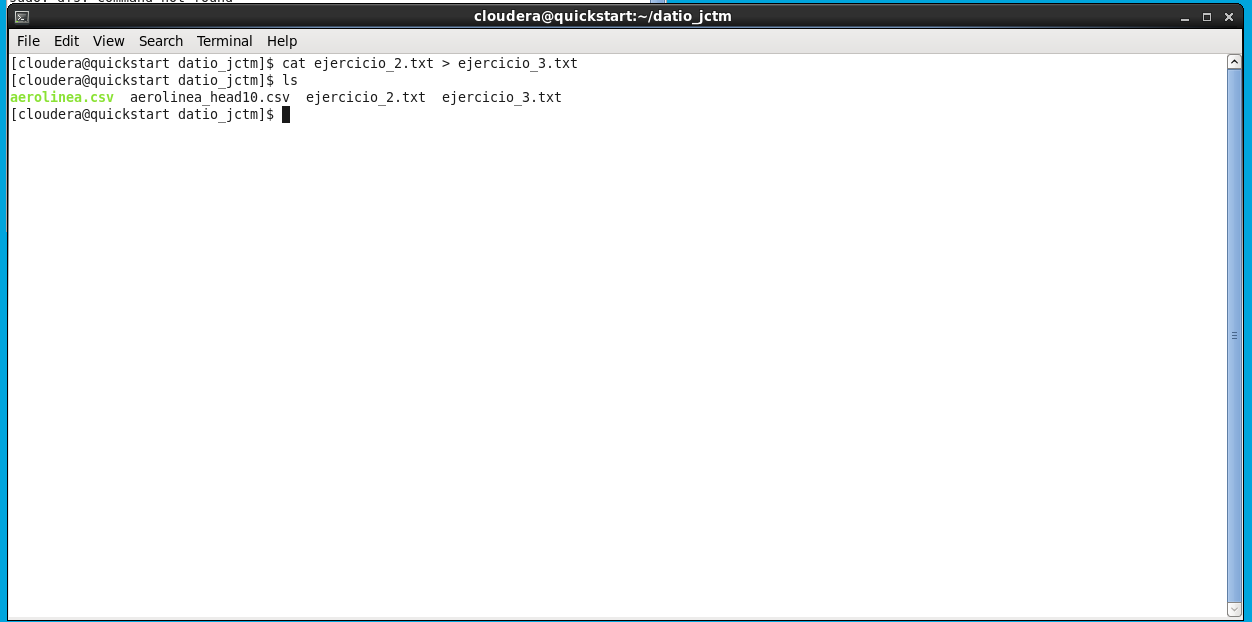
**http://www.linfo.org/pipes.html**

Respuesta : sed -n '1,25p; 25q' aerolinea.csv | tee **ejercicio\_2.txt**



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

Respuesta : cat ejercicio\_2.txt > ejercicio\_3.txt



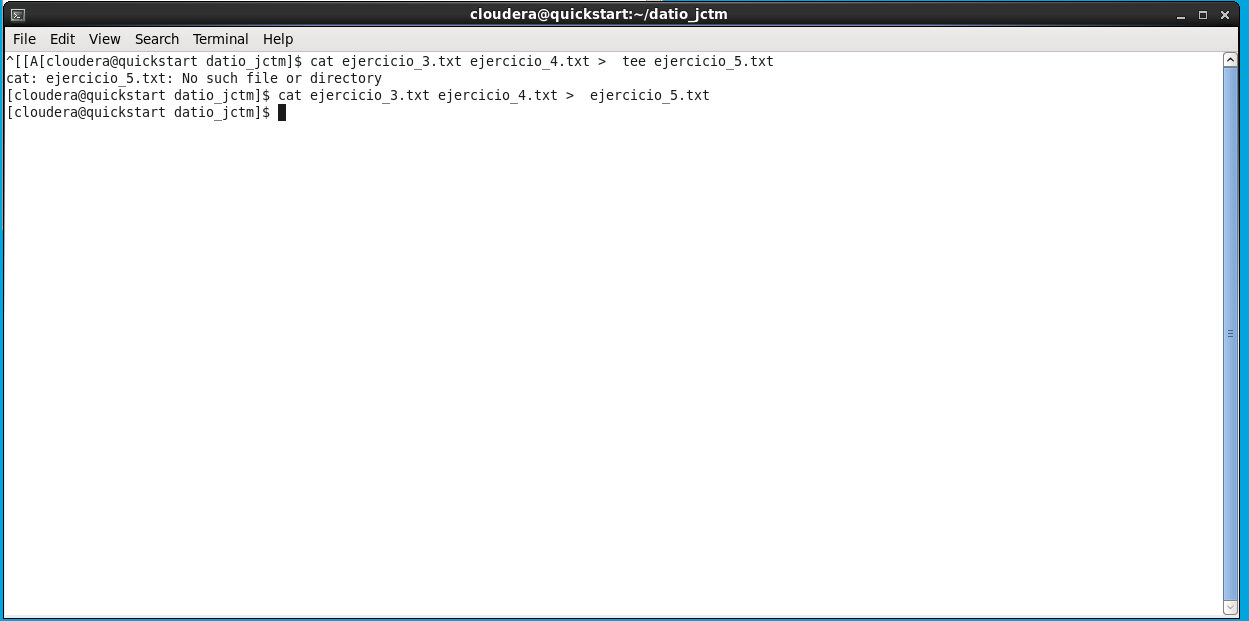
4.- Con algún comando en GNU/Linux tome las 25 últimas líneas del archivo aerolínea.csv **SIN** emplear el comando tail y guárdelo como **ejercicio\_4.txt**

**Resultado:** sed -n '123534947,123534972p; 123534972q' aerolinea.csv > ejercicio\_4.txt



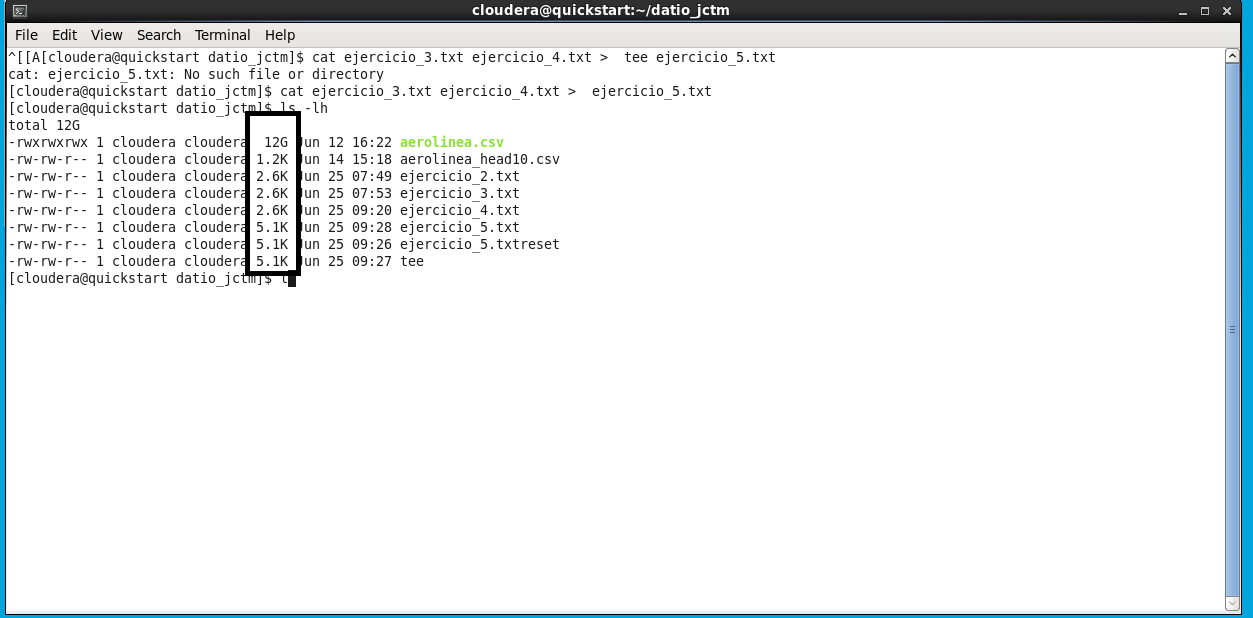
5.- Concatene los archivos **ejercicio\_3.txt** y **ejercicio\_4.txt** en un archivo **ejercicio\_5.txt** y en esa misma pantalla resultado muestre el contenido de **ejercicio\_5.txt**

**Resultado: cat ejercicio\_3.txt ejercicio\_4.txt > ejercicio\_5.txt**



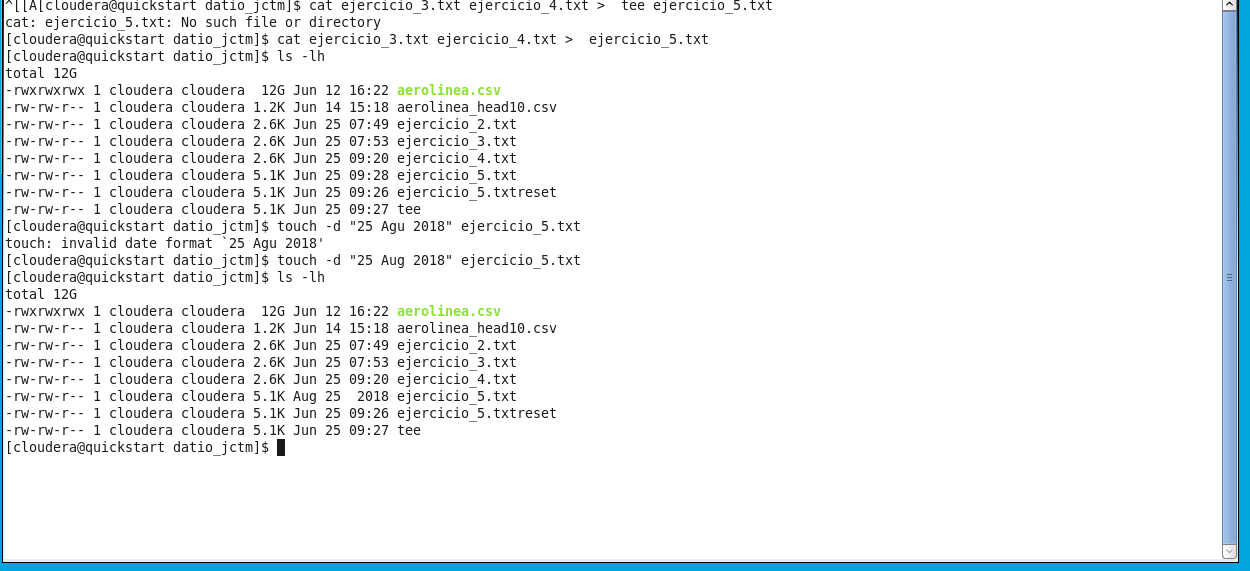
6.- Usando el comando **ls** y sus opciones, verifique el peso de **ejercicio\_5.txt,** señalando en la captura de pantalla dónde se encuentra éste.

**Resultado: ls -lh**

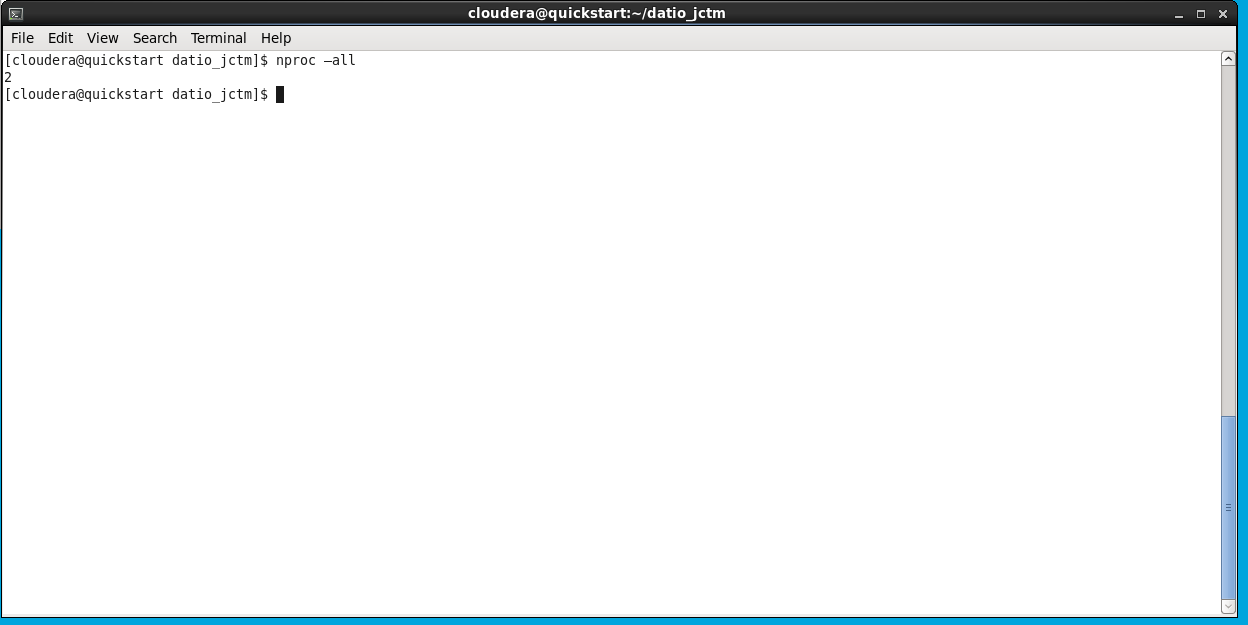


7.- Modifique la fecha de acceso de **ejercicio\_5.txt** al 25 de Agosto del 2018 y muestre en pantalla dónde se puede apreciar ese resultado.

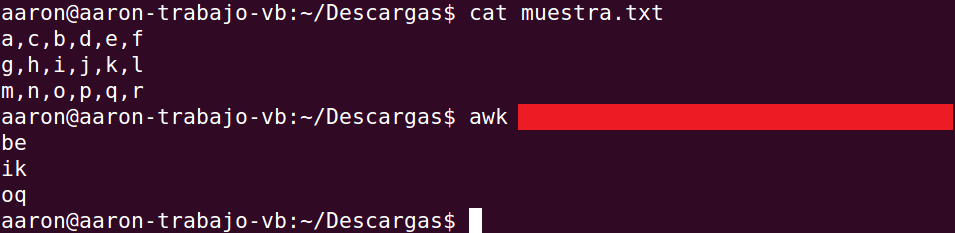
**Resultado:** touch -d "25 Aug 2018" ejercicio\_5.txt

  
  
8.- ¿Con cuál comando se puede averiguar el número de núcleos en un sistema GNU/Linux? Investigue y coloque el resultado, haciendo énfasis en el lugar donde se puede apreciar esa información.

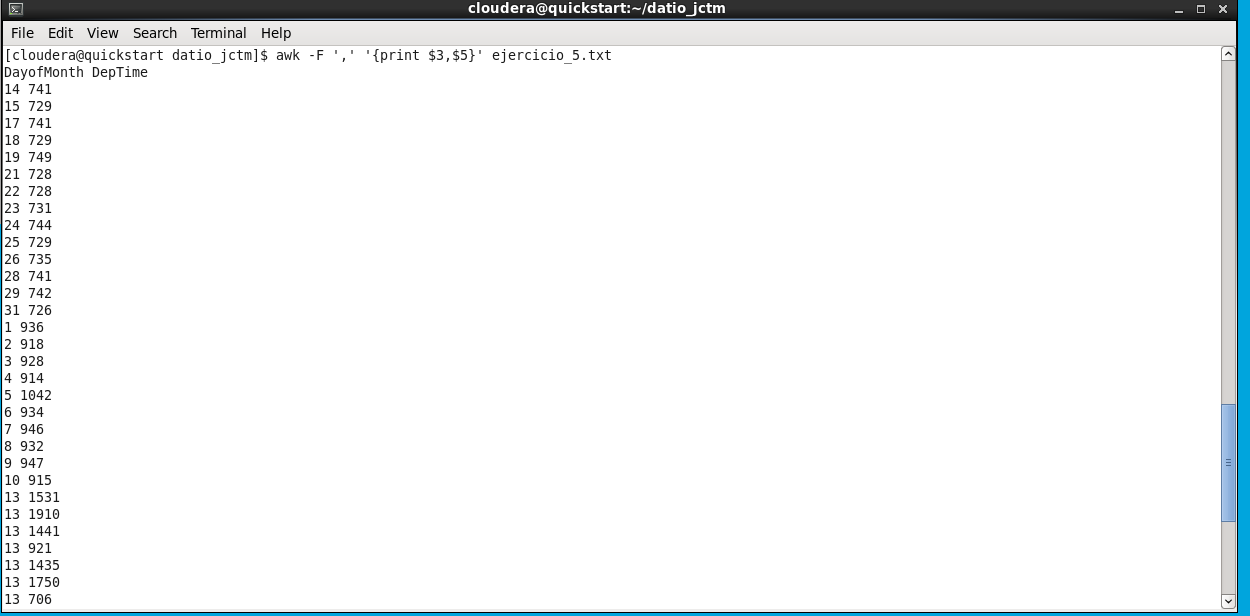
**Resultado:** nproc –all



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:

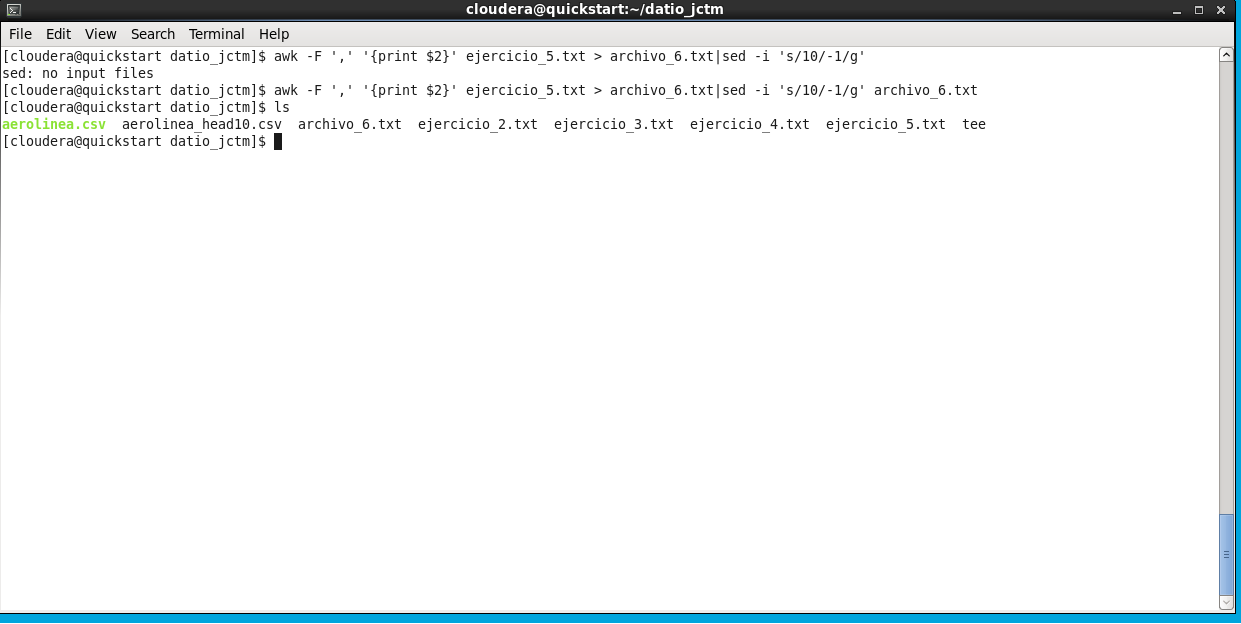


**Resultado: awk -F ',' '{print $3,$5}' ejercicio\_5.txt**



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.

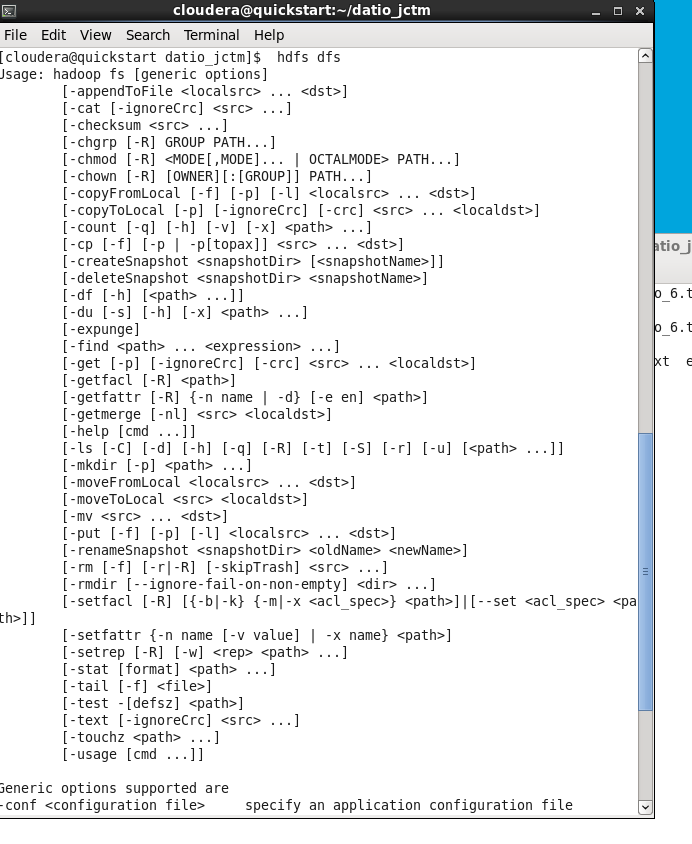
**Resultado: awk -F ',' '{print $2}' ejercicio\_5.txt > archivo\_6.txt|sed -i 's/10/-1/g' archivo\_6.txt**

****

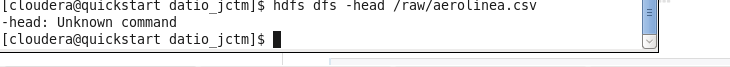
**SECCION 2. HDFS Y HIVE**

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

**hdfs dfs -head /raw/aerolínea.csv**

****

**En el HDFS no se tiene la function o commando head por lo cual la ejecucion de la linea “hdfs dfs -head /raw/aerolínea.csv” ejecuta un error**

****

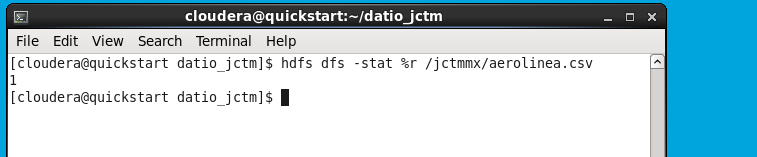
Con una captura muestre qué es lo que pasa y por medio de argumentos sólidos (una captura de pantalla con la evidencia, una fuente de consulta) por qué sucede esto.

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.cvs | wc –l**

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 /jctmmx/aerolínea.csv



14.- Tome como base el archivo **aerolínea.csv** del HDFS y almacene en el sistema local un archivo **ejercicio\_14.txt** que contenga las primeras 15 líneas sin usar el comando -tail del HDFS. Muestre ese contenido también.

hdfs dfs –cat /jctmmx/aerolínea.txt |head –n 15 | tee ejercicio\_14.txt

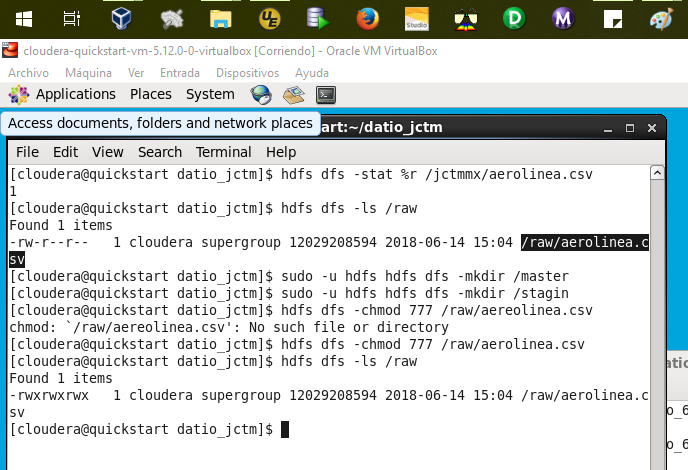
15.- Cree los directorios **master**  y **stagin** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso.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

hdfs dfs -ls /raw

sudo -u hdfs hdfs dfs -mkdir /master

sudo -u hdfs hdfs dfs -mkdir /stagin

hdfs dfs -ls /raw

  
  
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:

CREATE EXTERNAL TABLE tabla\_aerolinea(

Year STRING, Month STRING, DayofMonth STRING, DayOfWeek STRING, DepTime STRING, CRSDepTime STRING, ArrTime STRING, CRSArrTime STRING, UniqueCarrier STRING, FlightNum STRING, TailNum STRING, ActualElapsedTime STRING, CRSElapsedTime STRING, AirTime STRING, ArrDelay STRING, DepDelay STRING, Origin STRING, Dest STRING, Distance STRING,

TaxiIn STRING, TaxiOut STRING, Cancelled STRING, CancellationCode STRING, Diverted STRING,

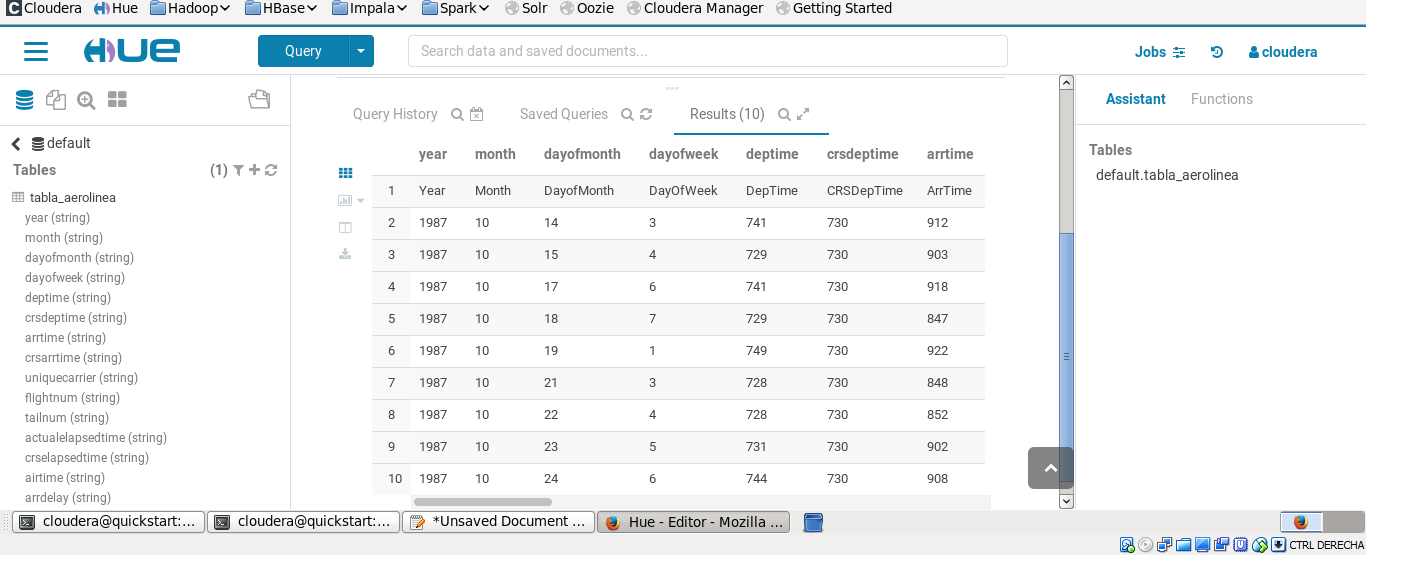
CarrierDelay STRING, WeatherDelay STRING,

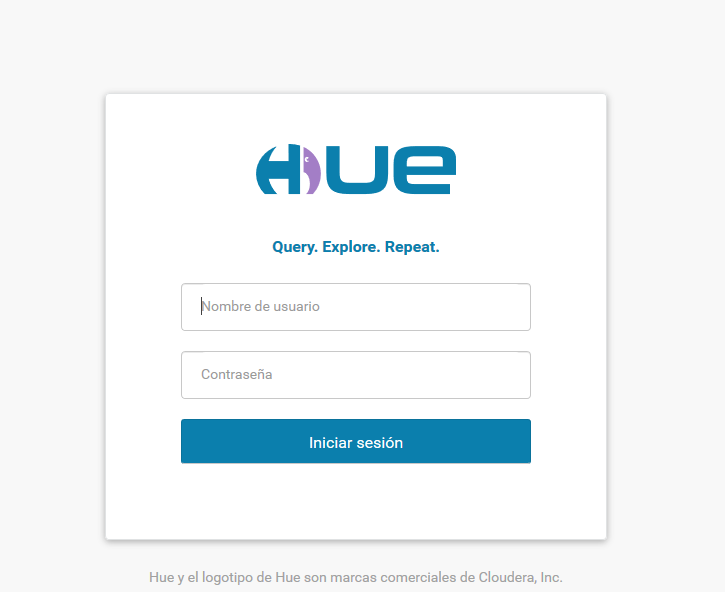
NASDelay STRING, SecurityDelay STRING, LateAircraftDelay STRING),

Adicional STRING

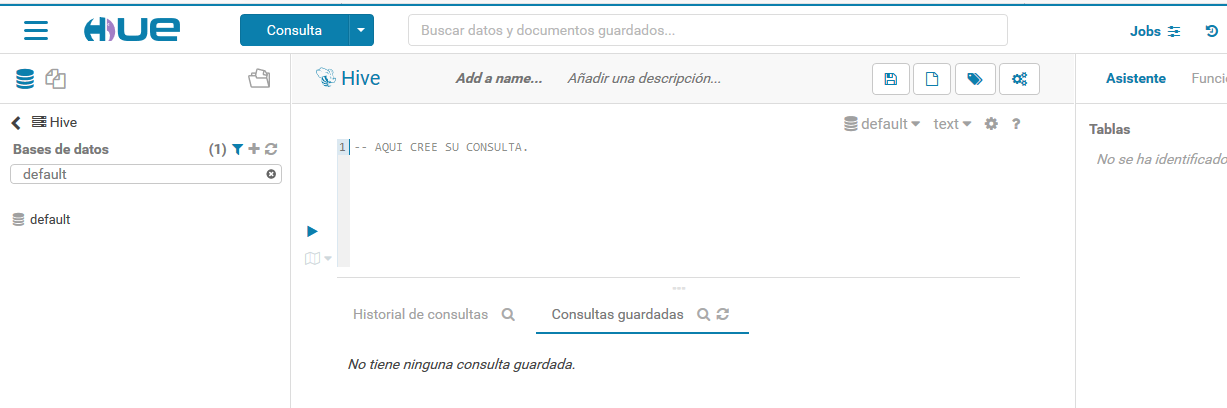
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE location '/raw'

TBLPROPERTIES ('SKIP.HEADER.LINE.COUNT'='1');





Recuerde que tanto el usuario como la contraseña es **cloudera**:

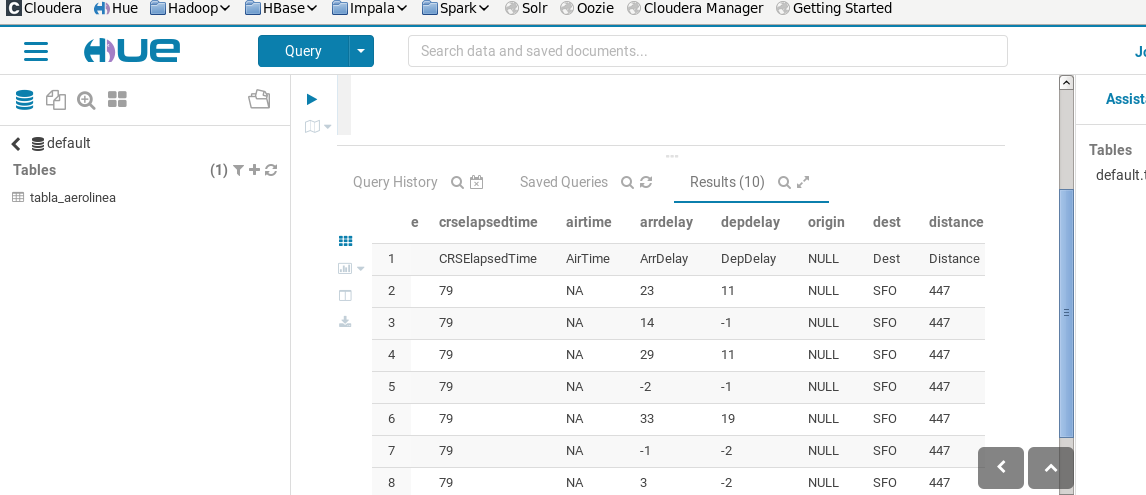


Entonces tome el siguiente código y cree una tabla en Hive:  
  
CREATE EXTERNAL TABLE tabla\_aerolinea(

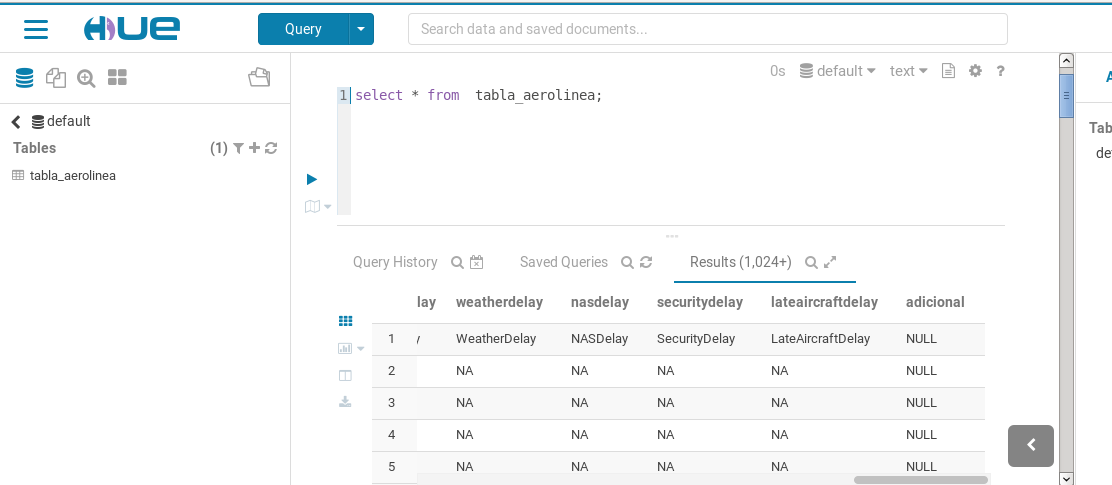
Year STRING,  
Month STRING,  
DayofMonth STRING,  
DayOfWeek STRING,  
DepTime STRING,  
CRSDepTime STRING,  
ArrTime STRING,  
CRSArrTime STRING,  
UniqueCarrier STRING,  
FlightNum STRING,  
TailNum STRING,  
ActualElapsedTime STRING,  
CRSElapsedTime STRING,  
AirTime STRING,  
ArrDelay STRING,  
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.

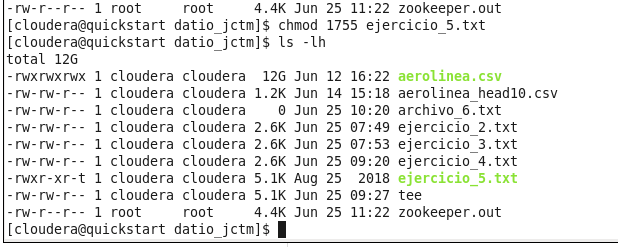


19.- En esta tabla anterior inserte un renglón a la tabla con todos los valores iguales a “NA” (tiene que investigar cómo añadir elementos a la tabla), y luego después de la inserción del elemento indague en qué parte del HDFS se ha guardado ese nuevo elemento.

**SECCIÓN 3. PREGUNTAS ABIERTAS**

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

Permiso de acceso a los fichos

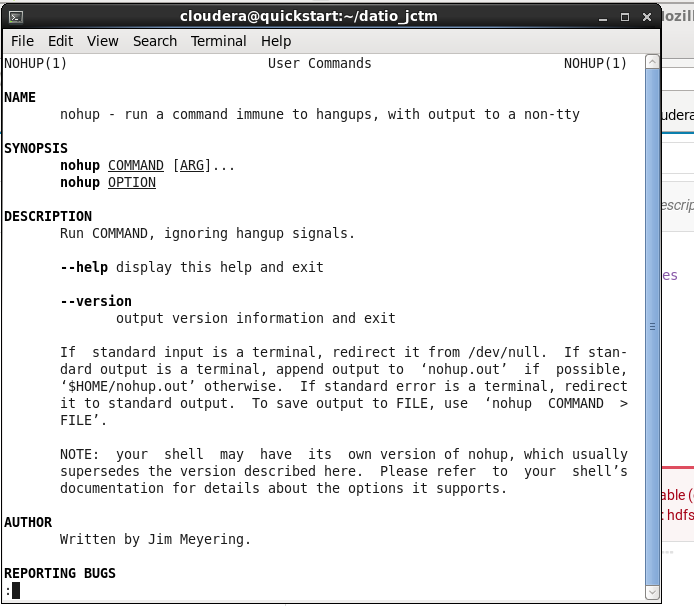
****

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

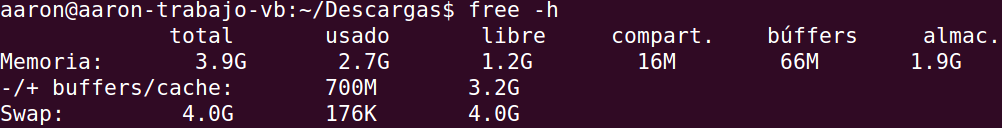
SQL para objetos que no son estructradas de las BDR, no requieren tablas estrutrudadas, hive administra e impala ejecta procesos nosql

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?

Comando que permite la ejecución de los procesos sin perder de vista el proceso



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

  
Indique el o los valores adecuados y por qué.

Es 3.9 G de los cuales usado tiene 2.7, el Swap es la parte de disco duro destinada para ser usada cono RAM

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 755 objetivo.txt

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

Cludera es la herramienta bigdata que incluye hadoop como manegador de archivos distribuids y

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

Unix/Linux

Ext, Ext2, Ext3, Swap

Windows

Fat32, FAT 16, NTFS

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

Es una función de serelzable y deserielizable de los archivos y se utiliza en en ambos, hive e impala

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

BigTable, es el sistema de Google de distribuido

BigQuery es el servicio web de Google para consultar

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

Data Lake, es la herramienta para realizar la ingesta de archivos

Data Warehouse, sistema estructurado que permite el proceso de info

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

Open GFS,

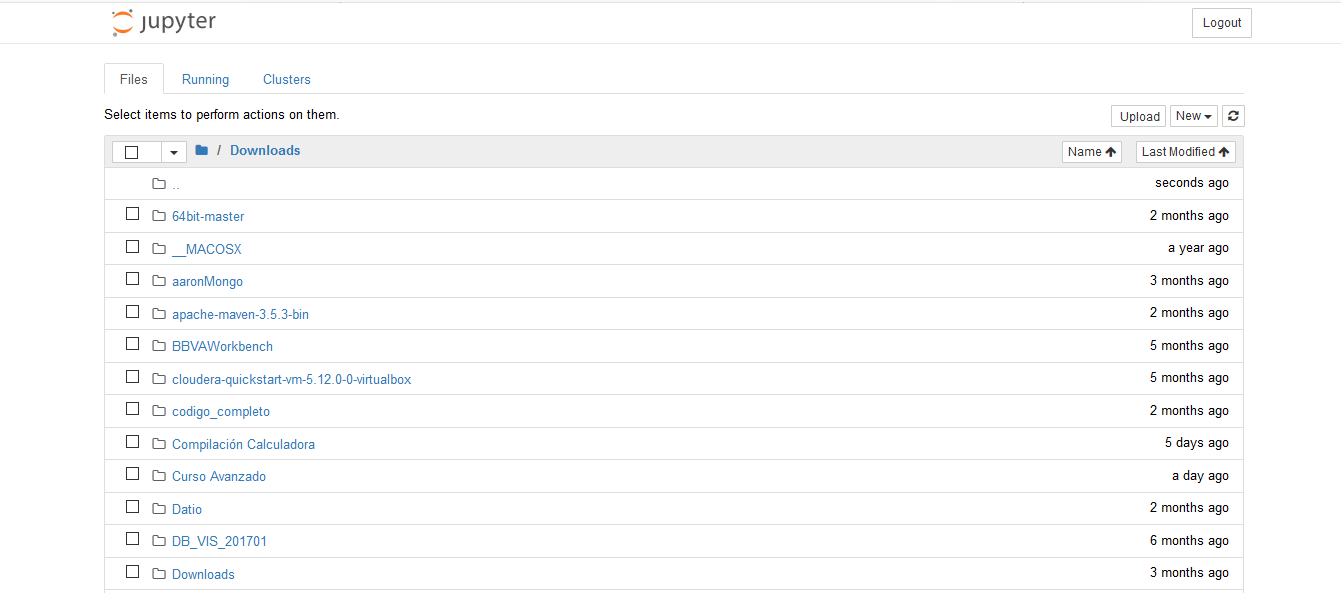
MooseFS.

**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:**



**sudo -u yum install –y** <http://dl.fedoraproject.org/pub/epel/7/x86_64/Packages/e/epel-release-7-11.noarch.rpm>

**sudo yum install -y python-pip python-devel python-virtualenv**

sudo yum groupinstall 'Development Tools'

**virtualenv jupyter-virtualenv**

source jupyter-virtualenv/bin/activate

pip install jupyter