

## Tarea de Vacaciones

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

El estudiante debe crear primero que nada un directorio dentro de su directorio local de Git llamado **TareaVacaciones** y dentro de éste crear una copia de esta tarea que lleve por nombre **TareaX y colocarla juntos con los resultados en formato PDF**, donde X es su nombre de usuario empleado en Github. Por ejemplo:

### TareaYoNoFui

Con respecto de los ejercicios, a menos que se indique lo contrario, todas las respuestas constarán del código o instrucción resultante acompañada de una captura de pantalla. Ejemplo:

-1.- Indique el comando que se emplea para listar archivos en GNU/Linux de manera simple:

Respuesta: **ls**

```
aaron@aaron-trabajo-vb:~/Descargas$ ls
apache-hive-2.3.3-bin.tar.gz      R-3.3.2.tar.gz
archivo.txt                       rattle_5.0.18.tar.gz
banner-principal2.png            Respaldo Usuaría Chile
core-site.xml                    RGtk2_2.20.33.tar.gz
db-derby-10.13.1.1-bin.tar.gz    rstudio-1.0.143-amd64.deb
file01.txt                       sas2txt.py
file02.txt                       scala-2.10.4.deb
file03.txt                       scala-2.12.1.tgz
```

En este tipo de preguntas de faltar alguno de los elementos señalados se considerará como errónea la respuesta y no se obtendrá el acierto.

Es menester mencionar que hay casos donde las preguntas son de tipo abierto, entonces en esos casos lo único que se pide adjuntar es tanto la respuesta como la(s) fuente(s). Ejemplo:

0.- ¿Cuál es el significado de la vida?

Respuesta: **42**

Fuente: <https://www.independent.co.uk/life-style/history/42-the-answer-to-life-the-universe-and-everything-2205734.html>

De nueva cuenta, si no existe al menos uno de estos dos elementos, la respuesta se considerará como inválida.

La fecha límite de entrega es el **Lunes 25 de Junio a las 15:00:00**, como se había mencionado con anterioridad el flujo de archivos se mantiene única y exclusivamente por Github, para ello se dejan los comandos a emplearse:

- **git pull** (para actualizar el repositorio)
- **git add .** (para indicar todos los elementos que se desean agregar al repositorio)
- **git commit -m "TareaVacaciones nombre\_usuario"** (para colocar un mensaje que distinga a esta subida de las de los demás usuarios)
- **git push origin master** (para efectuar los cambios)

Por cierto que en lo que se repara Git en Cloudera puede ocupar Git de Windows y de esta manera, ya que se sugirió la instalación de Guest Additions en VirtualBox, copiar los resultados al primer sistema.

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

Dicho lo anterior se les desea mucho éxito en la travesía, cualquier cosa no duden en preguntar...

## SECCION 1. GNU/LINUX

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

**R= sed 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>

**R= sed 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

**R=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**

```
R= sed -e :a -e '$q;N;26,$D;ba' 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**

```
R= cat ejercicio_3.txt ejercicio_4.txt | tee ejercicio_5.txt
```

```

applications Places System
cloudera@quickstart:~
Edit View Search Terminal Help
10,4,7,914,915,1003,1001,P5,1451,NA,49,46,NA,2,-1,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,5,1,1842,915,1129,1001,P5,1451,NA,47,46,NA,88,67,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,6,1,934,915,1024,1001,P5,1451,NA,50,46,NA,23,19,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,7,3,946,915,1037,1001,P5,1451,NA,51,46,NA,36,31,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,8,4,932,915,1033,1001,P5,1451,NA,61,46,NA,32,17,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,9,5,947,915,1036,1001,P5,1451,NA,49,46,NA,35,32,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
10,10,6,915,915,1022,1001,P5,1451,NA,67,46,NA,21,0,SFO,RNO,192,NA,NA,0,NA,0,NA,NA,NA,NA,NA
12,13,6,1910,1918,2017,1616,DL,1612,N927DA,67,66,38,1,0,ATL,CHS,259,5,24,0,,0,NA,NA,NA,NA,NA
12,13,6,1441,1445,1604,1622,DL,1613,N973DA,63,66,37,-18,-4,IND,ATL,432,8,10,0,,0,NA,NA,NA,NA,NA
12,13,6,921,830,1111,1008,DL,1616,N9070E,111,98,82,64,51,ATL,PBI,545,8,21,0,,0,51,0,13,0,0
12,13,6,1435,1440,1701,1704,DL,1618,N914DL,86,84,56,-3,MSY,ATL,425,20,10,0,,0,NA,NA,NA,NA,NA
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
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
12,13,6,1552,1528,1735,1718,DL,1620,N905DE,43,58,27,17,32,HSV,ATL,151,9,7,0,,0,0,0,0,0,17
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
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
12,13,6,840,843,1025,1021,DL,1624,N3388,105,98,53,4,-3,SLC,DEN,391,6,46,0,,0,NA,NA,NA,NA,NA
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
12,13,6,547,545,646,650,DL,1627,N621DL,59,65,38,-4,2,SAV,ATL,215,13,0,,0,NA,NA,NA,NA,NA
12,13,6,848,850,1024,1005,DL,1628,N920DL,156,135,108,19,-2,ATL,MCI,692,4,44,0,,0,0,19,0,0
12,13,6,936,936,1114,1119,DL,1630,N653DL,98,103,70,-8,ATL,RSW,514,24,0,,0,NA,NA,NA,NA,NA
12,13,6,657,609,904,749,DL,1631,N3743H,127,108,78,75,57,ATL,ATL,481,15,34,0,,0,57,18,0,0
12,13,6,1007,847,1149,1010,DL,1631,N909DA,162,143,122,99,08,ATL,IAH,669,8,32,0,,0,1,0,19,0,79
12,13,6,638,648,808,753,DL,1632,N604DL,90,73,56,-2,JAX,ATL,270,14,26,0,,0,0,0,15,0,0
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
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
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
12,13,6,1002,959,1204,1150,DL,1636,N646DL,122,111,71,14,3,ATL,IAH,53,6,45,0,,0,NA,NA,NA,NA,NA
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
12,13,6,655,708,856,856,DL,1638,N671DN,121,116,85,-5,-5,PBI,ATL,545,12,42,0,,0,NA,NA,NA,NA,NA
12,13,6,1251,1246,1446,1437,DL,1639,N646DL,115,117,89,9,11,ATL,ATL,53,13,13,0,,0,NA,NA,NA,NA,NA
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
dera@quickstart:~$
Cloudera Live: Welco... cloudera@quickstart:~

```

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.

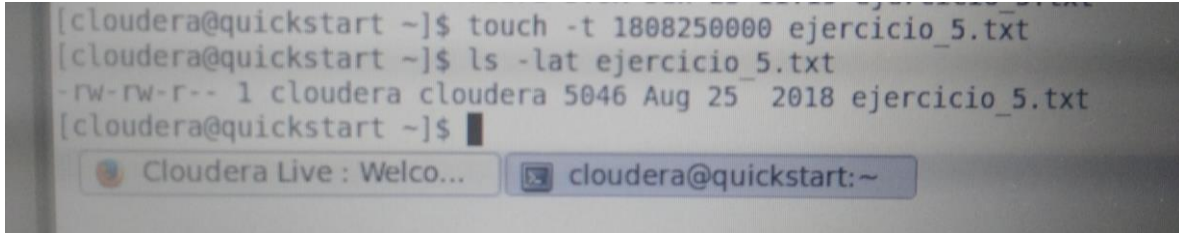
**R= ls -lah ejercicio\_5.txt**

2008,12,13,6,1110,1103,1413,1418,DL,1641,N908DL,123,135,104,-5,7,5  
[cloudera@quickstart ~]\$ ls -lah ejercicio\_5.txt  
-rw-rw-r-- 1 cloudera cloudera 5.0K Jun 25 11:19 ejercicio\_5.txt  
[cloudera@quickstart ~]\$

Cloudera Live : Welco... cloudera@quickstart:~

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.

R= `touch -t 1808250000 ejercicio_5.txt`

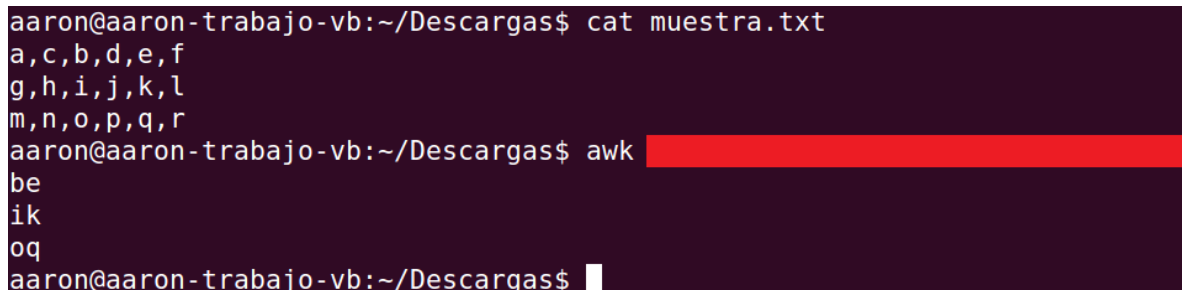


```
[cloudera@quickstart ~]$ touch -t 1808250000 ejercicio_5.txt
[cloudera@quickstart ~]$ ls -lat ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 5046 Aug 25 2018 ejercicio_5.txt
[cloudera@quickstart ~]$
```

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.

R= **NA**

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$
```

R= **AWK** utiliza un archivo o emisión de ordenes y un archivo o emisión de entrada. El primero indica como procesar al segundo. El archivo de entrada es por lo general texto con algún formato que puede ser un archivo o bien la salida de otro programa.

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.

R= **NA**

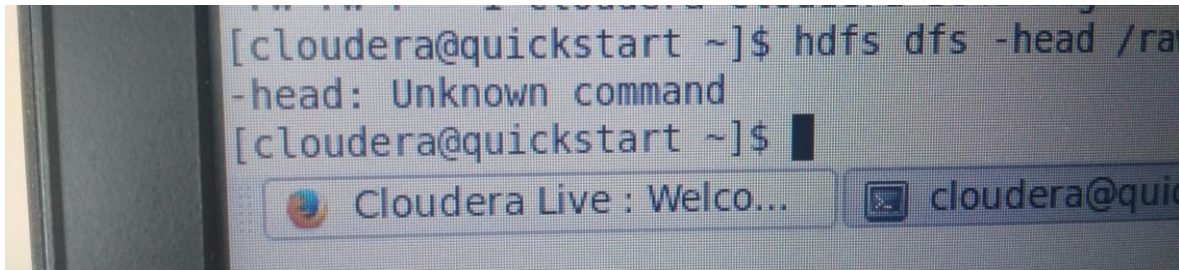
## SECCION 2. HDFS Y HIVE

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

**hdfs dfs -head /raw/aerolínea.csv**



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

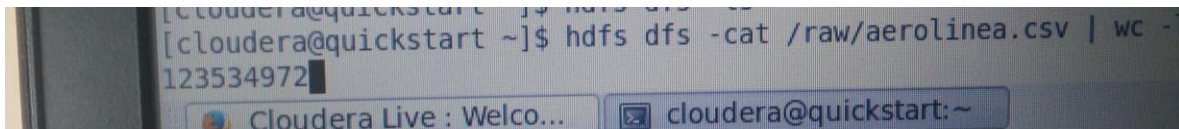


El hdfs no considera el `-head` como un comand.

Fuente: [https://www.tutorialspoint.com/es/hadoop/hadoop\\_command\\_reference.htm](https://www.tutorialspoint.com/es/hadoop/hadoop_command_reference.htm)

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

**R = `hdfs dfs -cat /raw/aerolinea.csv | wc -l`**

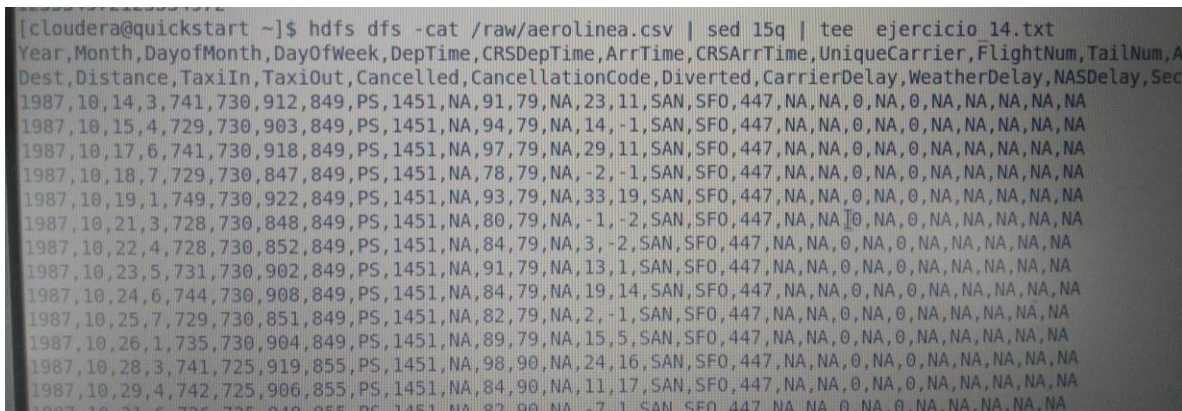


13.- Indague en la instrucción de HDFS para averiguar el factor de réplica del archivo **aerolínea.csv** y colóquelo aquí junto con captura del resultado.

**R= `hdfs getconf -confKey dfs.replication`**

**<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html>**

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.



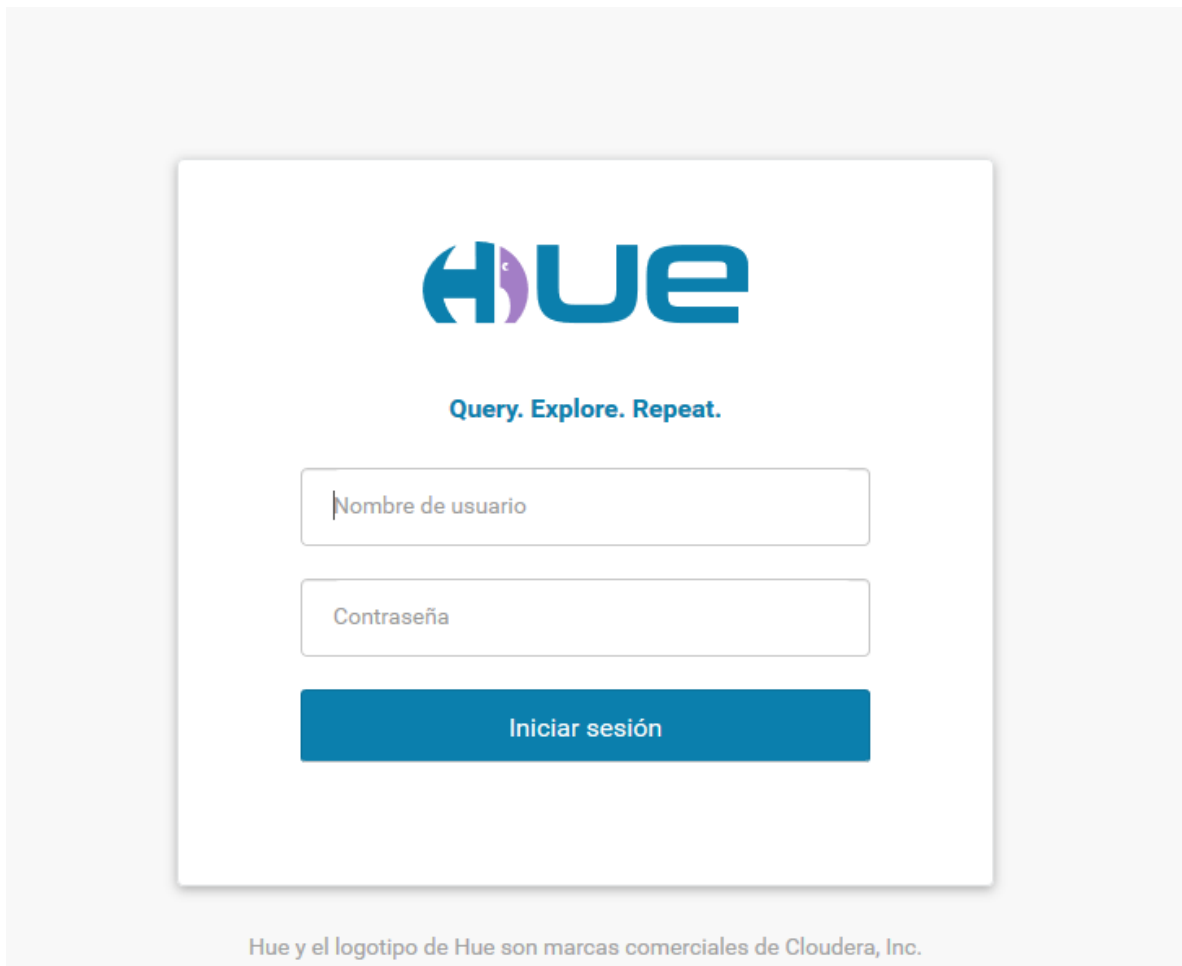
15.- Cree los directorios **master** y **staging** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso.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



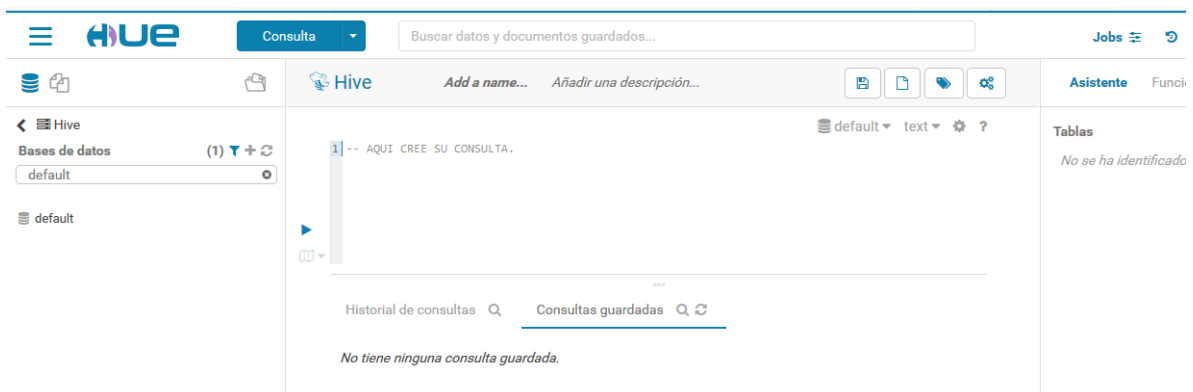
drwxrwxrwx	-	hdfs	supergroup	0	2017-07-19 05:34	/benchmarks
drwxr-xr-x	-	cloudera	supergroup	0	2018-06-05 21:28	/data
drwxr-xr-x	-	hbase	supergroup	0	2018-06-25 08:19	/hbase
drwxr-xr-x	-	cloudera	supergroup	0	2018-06-25 12:23	/master
drwxr-xr-x	-	cloudera	supergroup	0	2018-06-13 16:44	/raw
drwxr-xr-x	-	solr	solr	0	2017-07-19 05:37	/solr
drwxr-xr-x	-	hdfs	supergroup	0	2018-06-13 16:59	/staging
drwxrwxrwt	-	hdfs	supergroup	0	2018-05-30 04:29	/tmp
drwxr-xr-x	-	hdfs	supergroup	0	2018-06-05 21:05	/user
drwxr-xr-x	-	hdfs	supergroup	0	2017-07-19 05:36	/var

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';
```

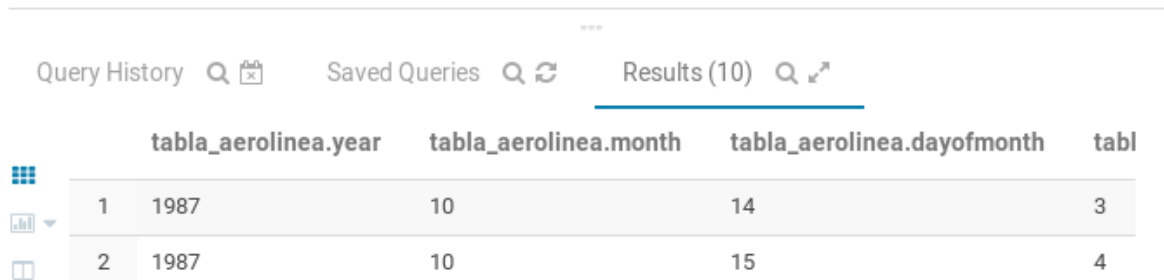
En el código anterior **NO** existe una forma de omitir los encabezados por lo que es su deber encontrar esa manera, incluirla en el código y crear la tabla.



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.

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

```
1 SELECT *
2 FROM tabla_aerolinea
3 LIMIT 10;
4
```



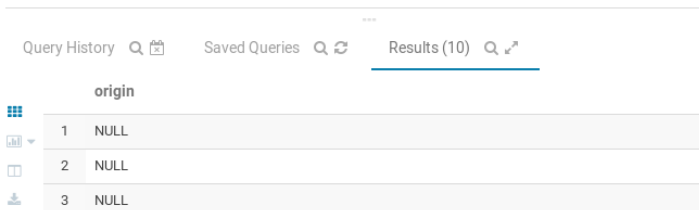
	tabla_aerolinea.year	tabla_aerolinea.month	tabla_aerolinea.dayofmonth	tbl
1	1987	10	14	3
2	1987	10	15	4

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 muéstrela.

```
17 DepDelay STRING,
18 Origin INT,
19 Dest STRING
```

El dato de Origin es de tipo alfanumérico entonces columna Origin ahora se informa en Nulos

```
1 SELECT ORIGIN
2 FROM tabla_aerolinea
3 LIMIT 10;
```



	origin
1	NULL
2	NULL
3	NULL

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.

R= la tabla se creó pero **Adicional** viene en nulo.

```
32 ROW FORMAT DELIMITED
33 FIELDS TERMINATED BY ','
34 STORED AS TEXTFILE
35 location '/raw'
36 tblproperties ("skip.header.line.count"="1");
37
```

✓ Success.

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.

R= NA

### SECCIÓN 3. PREGUNTAS ABIERTAS

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

R=Sticky bit

El Sticky bit se utiliza para permitir que cualquiera pueda escribir y modificar sobre un archivo o directorio, pero que solo su propietario o root pueda eliminarlo.

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

R= Las bases de datos NoSQL, también llamadas No Solo SQL, son un enfoque hacia la gestión de datos y el diseño de base de datos que es útil para grandes conjuntos de datos distribuidos.

Hive implementa una variante al SQL, llamada HQL (Hive QL). Esta variante no soporta la especificación SQL-92 completa, pero sí una gran parte. Realmente, cuando se ejecutan consultas HQL, Hive convierte la consulta HQL a un trabajo MapReduce que es ejecutado para obtener los datos. La finalidad de esto es permitir a usuarios que no cuenten con experiencia en programación de algoritmos MapReduce (que suelen ser bastante laboriosos), pero que sí cuenten con conocimientos de SQL, poder consultar datos.

Impala está dirigido a los analistas y científicos de datos para realizar análisis en los datos almacenados en Hadoop a través de herramientas de SQL o business intelligence.

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?

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

```
aaron@aaron-trabajo-vb:~/Descargas$ free -h
```

	total	usado	libre	compart.	buffers	almac.
Memoria:	3.9G	2.7G	1.2G	16M	66M	1.9G
-/+ buffers/cache:		700M	3.2G			
Swap:	4.0G	176K	4.0G			

Indique el o los valores adecuados y por qué.

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

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

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

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

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

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

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

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

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

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

#### SECCIÓN 4. ESPECIAL

31.- Instale Jupyter en Cloudera, para ello puede basarse en la siguiente liga:

<https://medium.com/@vando/install-jupyter-notebook-on-centos-7-1d596abf08da>

Es importante señalar que para continuar el curso es imprescindible esta herramienta y no existirán pausas para su instalación durante las sesiones, motivo por el cual es menester llevar a cabo esta operación aunque solamente valga 1 crédito. Para validar este ejercicio se requiere una captura de pantalla del menú principal, algo así:

Files Running Clusters

Select items to perform actions on them.

Upload New

<input type="checkbox"/>		/ Downloads	Name	Last Modified
		...		seconds ago
<input type="checkbox"/>		64bit-master		2 months ago
<input type="checkbox"/>		_MACOSX		a year ago
<input type="checkbox"/>		aaronMongo		3 months ago
<input type="checkbox"/>		apache-maven-3.5.3-bin		2 months ago
<input type="checkbox"/>		BBVAWorkbench		5 months ago
<input type="checkbox"/>		cloudera-quickstart-vm-5.12.0-0-virtualbox		5 months ago
<input type="checkbox"/>		codigo_completo		2 months ago
<input type="checkbox"/>		Compilación Calculadora		5 days ago
<input type="checkbox"/>		Curso Avanzado		a day ago
<input type="checkbox"/>		Datio		2 months ago
<input type="checkbox"/>		DB_VIS_201701		6 months ago
<input type="checkbox"/>		Downloads		3 months ago