

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.

\$ sed -n 1,10p aerolinea.csv

```

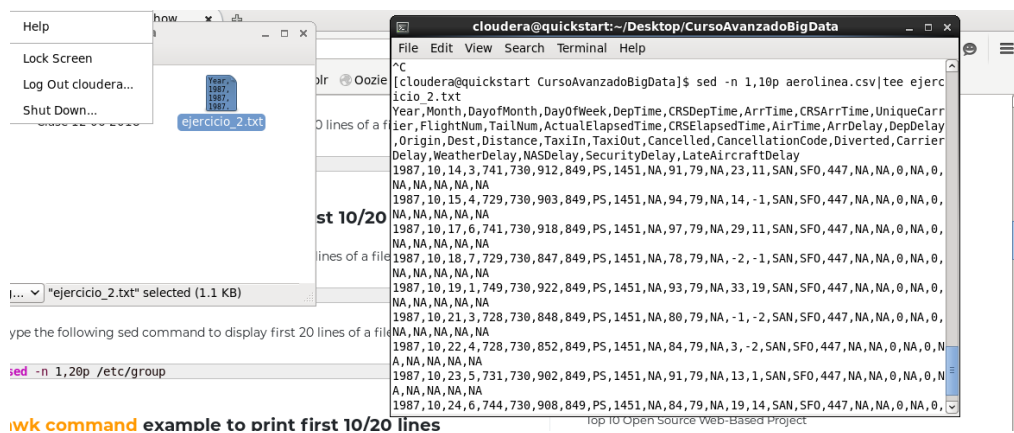
cloudera@quickstart:~/Desktop/CursoAvanzadoBigData
File Edit View Search Terminal Help
sed: -e expression #1, char 2: missing command
[cloudera@quickstart CursoAvanzadoBigData]$ sed -n 1,10p aerolinea.csv
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,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

```

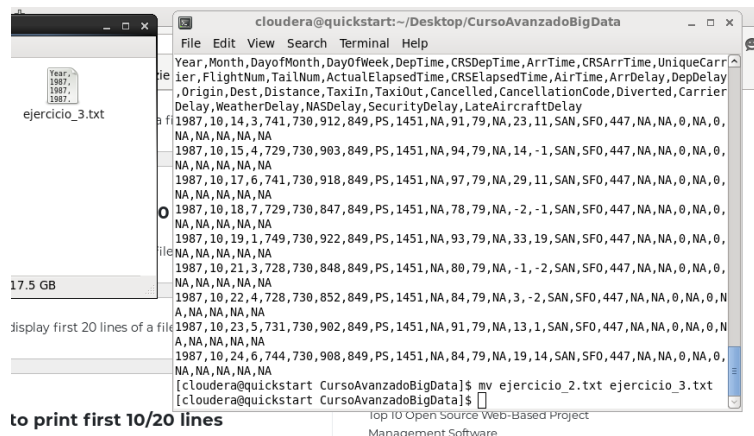
2.- Como ya se ha visto, utilizar el redireccionamiento destructivo (>) implica almacenar típicamente algún contenido en un archivo (ej. **echo "contenido" > archivo**). Pero lo cierto es que con este comando no se apreciará en pantalla lo que se desea almacenar en dicho archivo, por ello es que se necesita 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 1,10p aerolinea.csv | tee ejercicio_2.txt



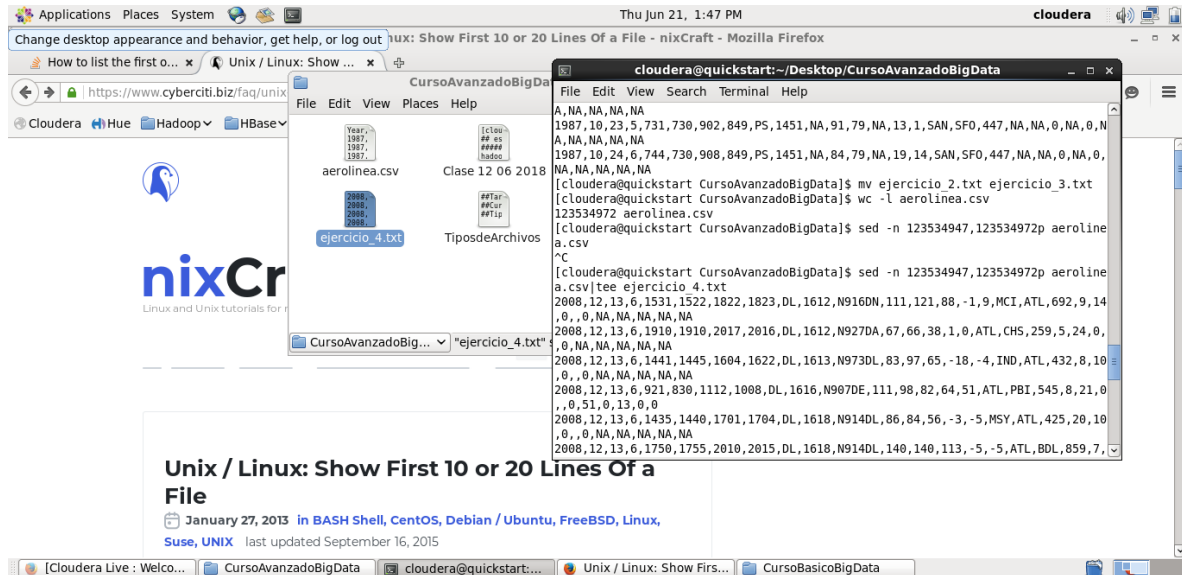
3.- Cambie el nombre del archivo **ejercicio_2.txt** a **ejercicio_3.txt** SIN usar el comando **rename**
mv ejercicio_2.txt ejercicio_3.txt



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**

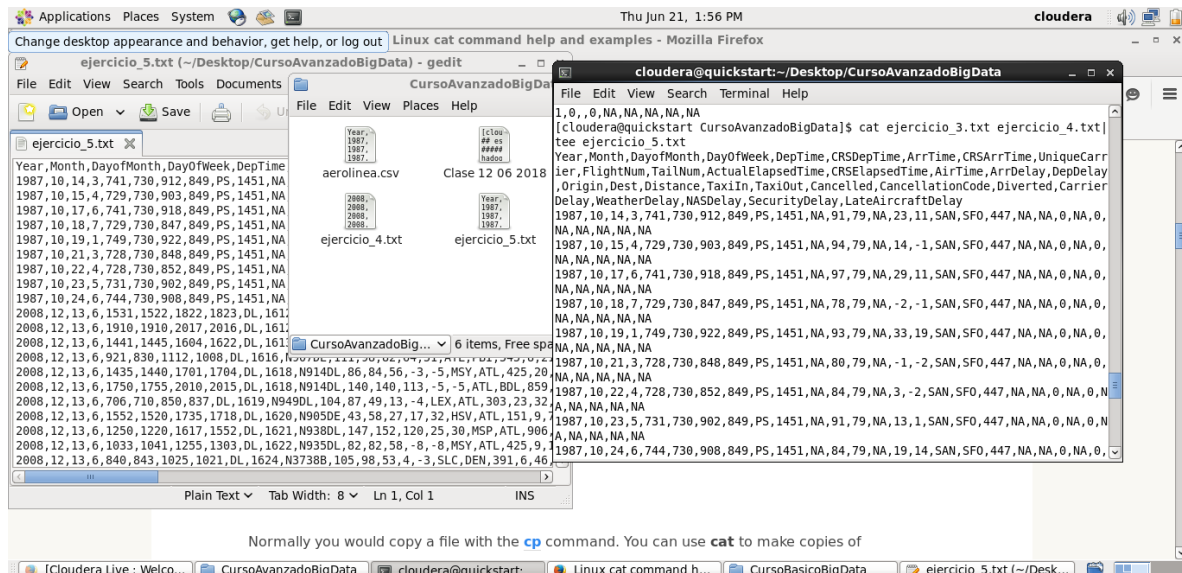
```
$wc -l aerolínea.csv
```

```
$sed -n 123534947, 1235372p | tee 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**

```
$ cat ejercicio_3.txt ejercicio_4.txt | tee 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.

```
$ ls -lh ejercicio_5.txt
```



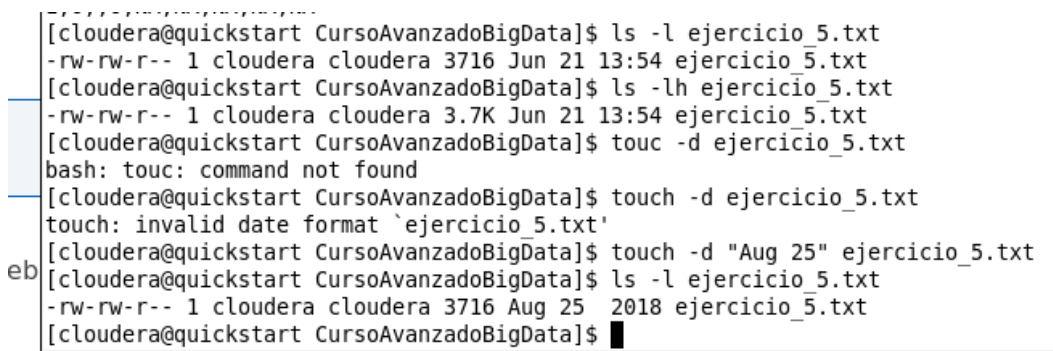
```
[cloudera@quickstart CursoAvanzadoBigData]$ ls -lh ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3.7K Jun 21 13:54 ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$
```

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.

```
$ ls -l ejercicio_5.txt
```

```
$ touch -d "Aug 25" ejercicio_5.txt
```

```
$ ls -l ejercicio_5.txt
```



```
[cloudera@quickstart CursoAvanzadoBigData]$ ls -l ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3716 Jun 21 13:54 ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$ ls -lh ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3.7K Jun 21 13:54 ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$ touch -d ejercicio_5.txt
bash: touch: command not found
[cloudera@quickstart CursoAvanzadoBigData]$ touch -d ejercicio_5.txt
touch: invalid date format `ejercicio_5.txt'
[cloudera@quickstart CursoAvanzadoBigData]$ touch -d "Aug 25" ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$ ls -l ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3716 Aug 25 2018 ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$
```

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.

El método más sencillo para saber el número de procesadores presentes en una máquina es

\$ nproc -all

Otra forma de obtener el mismo resultado, el cual también nos permite obtener información adicional sobre nuestro procesador es el comando **lscpu**:

\$ lscpu | grep 'CPU(S)'

Podemos encontrar numerosos detalles adicionales sobre nuestros procesadores en /proc/cpuinfo, entre ellos el modelo de CPU y el número de núcleos o cores que tiene:

\$ cat /proc/cpuinfo | grep "model name"

\$ cat /proc/cpuinfo | grep "cpu cores"

```
touch: invalid date format `ejercicio_5.txt'
[cloudera@quickstart CursoAvanzadoBigData]$ touch -d "Aug 25" ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$ ls -l ejercicio_5.txt
-rw-rw-r-- 1 cloudera cloudera 3716 Aug 25 2018 ejercicio_5.txt
[cloudera@quickstart CursoAvanzadoBigData]$ nproc --all
1
[cloudera@quickstart CursoAvanzadoBigData]$ lscpu|grep 'CPU(s)'
CPU(s): 1
On-line CPU(s) list: 0
NUMA node0 CPU(s): 0
[cloudera@quickstart CursoAvanzadoBigData]$ cat /proc/cpuinfo|grep "model name"
model name      : Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz
[cloudera@quickstart CursoAvanzadoBigData]$ cat /proc/cpuinfo|grep "cpu_cores"
[cloudera@quickstart CursoAvanzadoBigData]$ cat /proc/cpuinfo|grep "cpu cores"
cpu cores       : 1
[cloudera@quickstart CursoAvanzadoBigData]$
```

Fuente: www.daniloaz.com/es/como-saber-cuantos-procesadores-y-nucleos-tiene-una-maquina-linux/

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 'NR==3{print $3,$5}' muestra.txt
be
ik
oq
aaron@aaron-trabajo-vb:~/Descargas$
```

El comando awk es un método muy potente que permite procesar o analizar archivos de texto que están organizados por líneas (filas) y columnas. El formato básico del comando awk es el siguiente:

awk 'condición {acción}' archivo-entrada > archivo-salida

\$ awk '{print}' ejercicio_5.txt

\$ awk -F',' '{print \$1}' ejercicio_5.txt

\$ awk -F',' '{print \$3,\$5}' ejercicio_5.txt

SECCION 2. HDFS Y HIVE

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

```
hdfs dfs -head /raw/aerolínea.csv
```

El archive aun no se encuentra en el hdfs en este momento. Un head puede ser replicado usando un pipe en la salida de un cat por ejemplo hadop fs –cat a travez de la línea de comandos.

```
hadoop fs -cat /path/to/file | head
```

esto es por eficiencia, ya que con el commando tail resulta mas eficiente realizarlo de la siguiente manera, ya que el comando fs –tail trabaja en el ultimo kilobyte, por lo que hadoop puede encontrar de manera más eficiente el último bloque y saltarse la última posición del kilobyte.

Fuente: <https://stackoverflow.com/questions/19778137/why-is-there-no-hadoop-fs-head-shell-command>

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.

Si falla la conexión a hdfs ejecutar los siguientes comandos:

```
$ sudo service hadoop-hdfs-datanode start
$ sudo service hadoop-hdfs-namenode start
$ sudo -u hdfs hdfs dfsadmin -safemode leave
$ sudo -u hdfs Hadoop fs -mkdir /user/Tarea
```

<https://drive.google.com/drive/folders/1aYmpS2X8Ow-fm8l0NbaMD9w8OWHqRODW?usp=sharing>

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

Es importante mencionar que para los ejercicios de hdfs se tomo una muestra de “aerolines.csv” por razones de espacio.

```
Hdfs dfs -cat /benchmarks/archivodfs.txt | wc -l
```

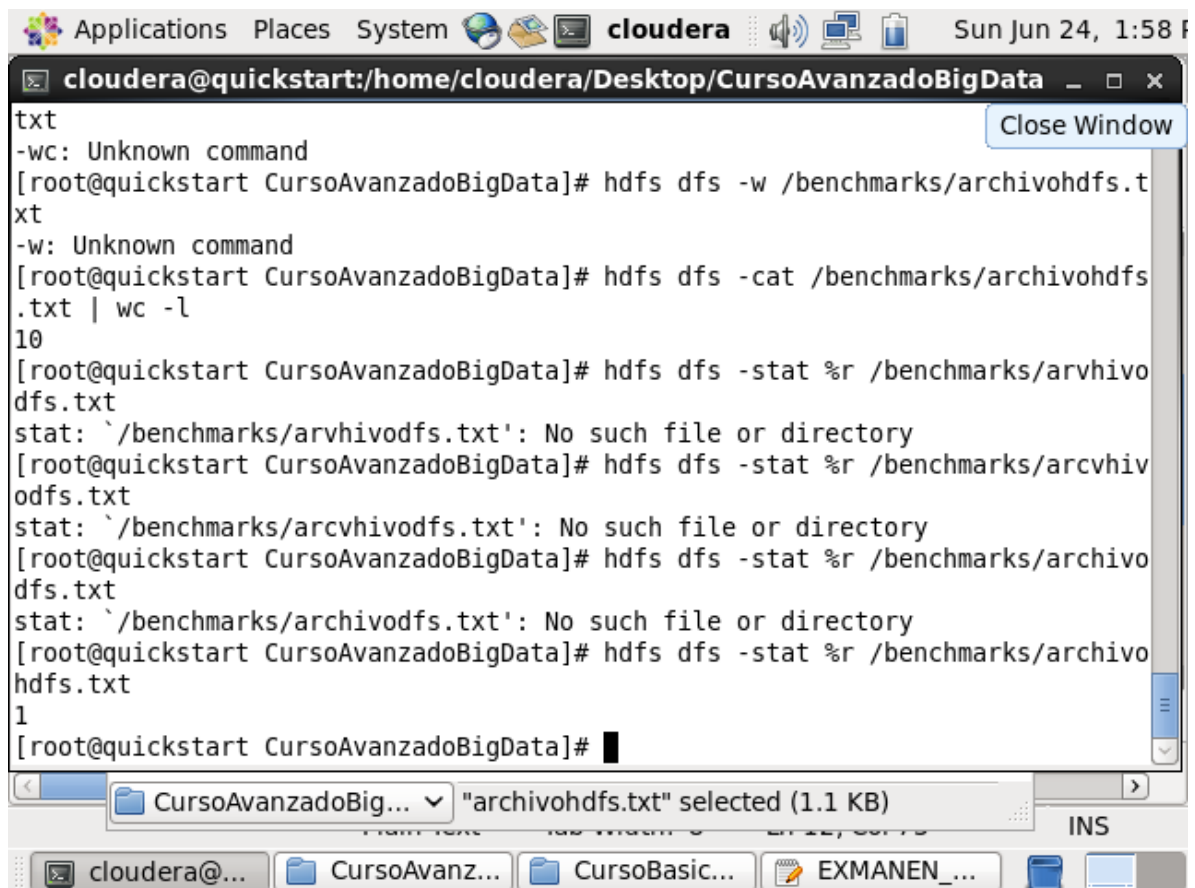

The screenshot shows a terminal window titled "cloudera@quickstart:/home/quickstart/cursoavanzadobigdata". The terminal displays the following commands and output:

```
drwxr-xr-x - solr solr 17-07-19 05:37 /solr
drwxrwxrwt - hdfs supergroup 0 2018-03-12 10:45 /tmp
drwxr-xr-x - hdfs supergroup 0 2018-06-21 16:36 /user
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /var
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -ls /benchmarks
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -put archivohdfs.txt /benchmarks
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -ls /benchmarks/
Found 1 items
-rw-r--r-- 1 root supergroup 1153 2018-06-24 13:45 /benchmarks/archivohdfs.txt
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -wc /benchmarks/archivohdfs.txt
-wc: Unknown command
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -w /benchmarks/archivohdfs.txt
-w: Unknown command
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -cat /benchmarks/archivohdfs.txt | wc -l
10
[root@quickstart CursoAvanzadoBigData]#
```

Below the terminal window, a file manager shows "CursoAvanzadoBig..." with "archivohdfs.txt" selected (1.1 KB). The taskbar at the bottom includes icons for "cloudera@...", "CursoAvanz...", "CursoBasic...", and "EXMANEN_...".

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 /benchmarks/archivohdfs.txt

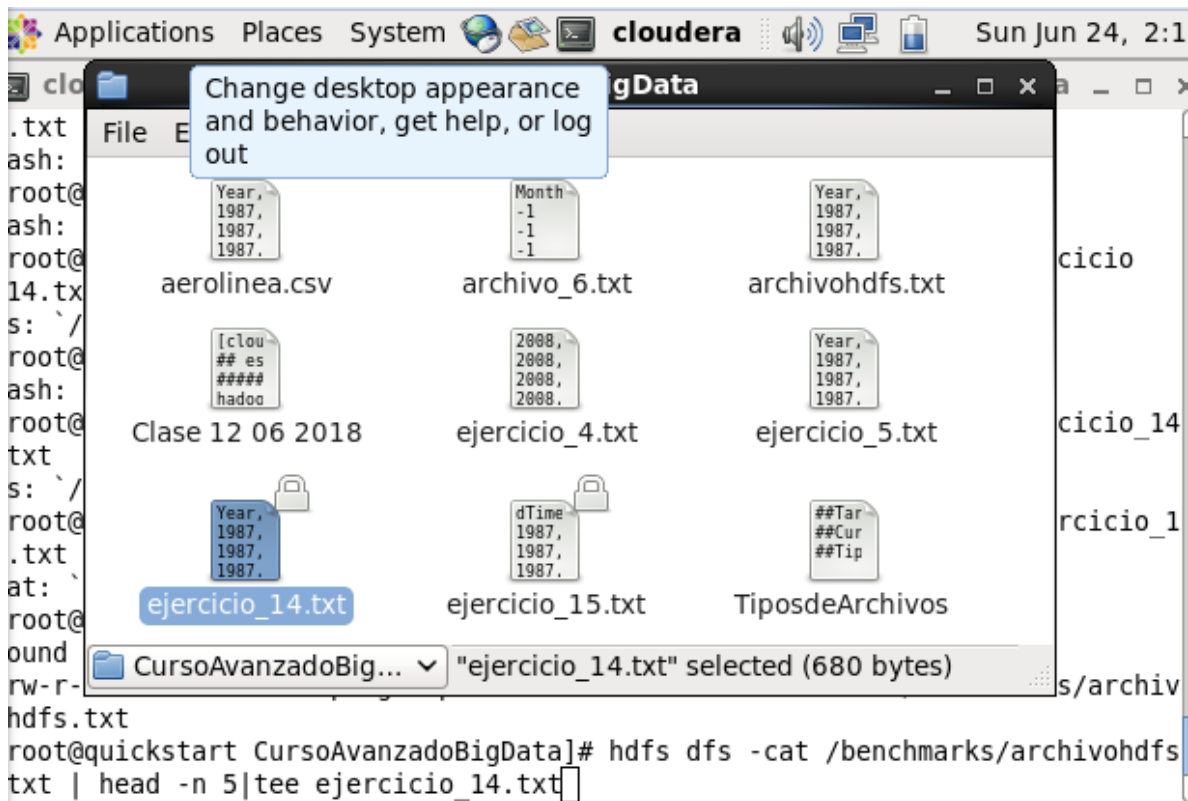


The screenshot shows a terminal window titled "cloudera@quickstart:/home/cloudera/Desktop/CursoAvanzadoBigData". The terminal output shows several failed HDFS commands due to file path errors. The user attempts to write, cat, and stat files in the path "/benchmarks/archivohdfs.txt", but the system reports "No such file or directory". The terminal ends with a prompt for a new command. Below the terminal, a file manager window shows "CursoAvanzadoBigData" selected, with a sub-window indicating "archivohdfs.txt" is selected (1.1 KB). The taskbar at the bottom shows several open applications, including "cloudera@...", "CursoAvanz...", "CursoBasic...", and "EXMANEN_...".

```
txt
-wc: Unknown command
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -w /benchmarks/archivohdfs.txt
-w: Unknown command
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -cat /benchmarks/archivohdfs.txt | wc -l
10
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -stat %r /benchmarks/arvhivodfs.txt
stat: `/benchmarks/arvhivodfs.txt': No such file or directory
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -stat %r /benchmarks/arcvhivodfs.txt
stat: `/benchmarks/arcvhivodfs.txt': No such file or directory
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -stat %r /benchmarks/archivodfs.txt
stat: `/benchmarks/archivodfs.txt': No such file or directory
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -stat %r /benchmarks/archivohdfs.txt
1
[root@quickstart CursoAvanzadoBigData]#
```

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 /benchmarks/archivohdfs.txt | head -n 15 | tee ejercicio_14.txt

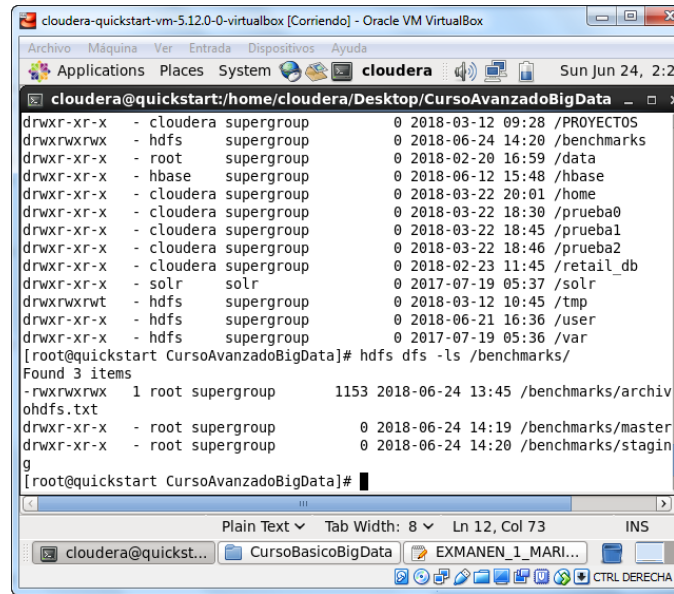


15.- Cree los directorios **master** y **staging** en el directorio raíz del HDFS y además al archivo aerolínea.csv que está en raw cámbiele los permisos de tal manera que el propietario tenga todas las facilidades sobre él, el grupo sólo pueda leer y escribir y cualquier otro no tenga ningún permiso.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

```
$hdfs dfs -mkdir /bechmarks/master
```

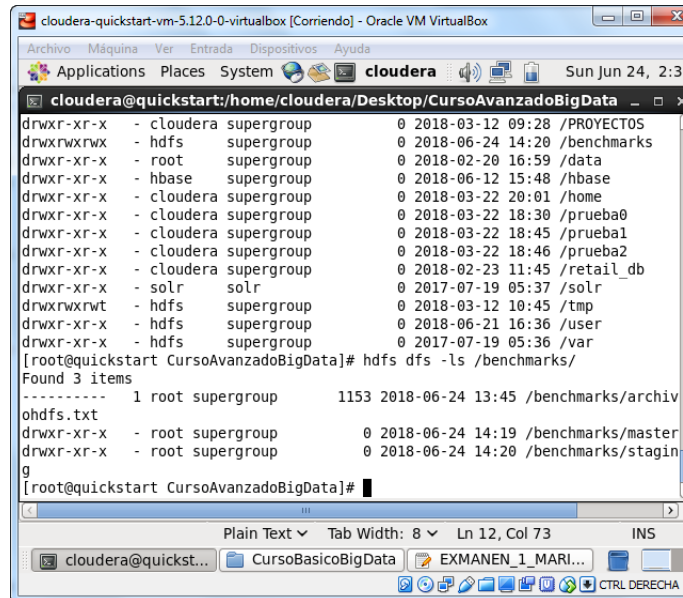
```
$hdfs dfs -mkdir /bechmarks/stagin
```

\$hdfs dfs -chmod 777 /bechmarks/archivohdfs.txt



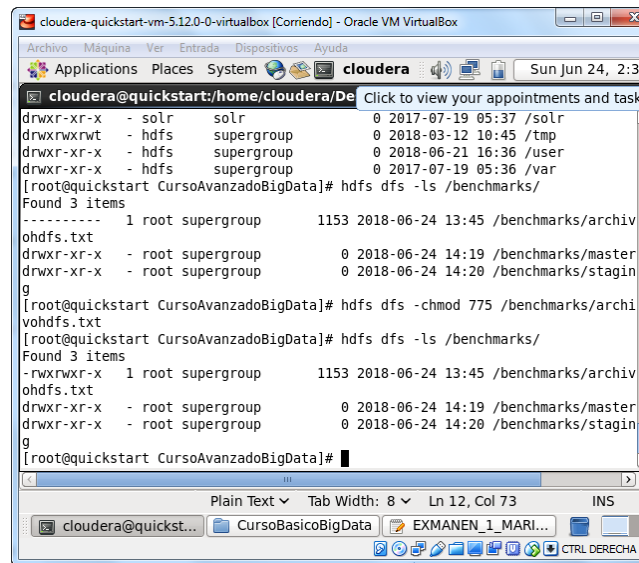
```
cloudera@quickstart:/home/cloudera/Desktop/CursoAvanzadoBigData$ hdfs dfs -ls /benchmarks/
Found 3 items
-rwxrwxrwx 1 root supergroup 1153 2018-06-24 13:45 /benchmarks/archivohdfs.txt
drwxr-xr-x - root supergroup 0 2018-06-24 14:19 /benchmarks/master
drwxr-xr-x - root supergroup 0 2018-06-24 14:20 /benchmarks/staging
```

\$ hdfs dfs -chmod 000 /benchmarks/archivohdfs.txt



```
cloudera@quickstart:/home/cloudera/Desktop/CursoAvanzadoBigData$ hdfs dfs -ls /benchmarks/
Found 3 items
----- 1 root supergroup 1153 2018-06-24 13:45 /benchmarks/archivohdfs.txt
drwxr-xr-x - root supergroup 0 2018-06-24 14:19 /benchmarks/master
drwxr-xr-x - root supergroup 0 2018-06-24 14:20 /benchmarks/staging
```

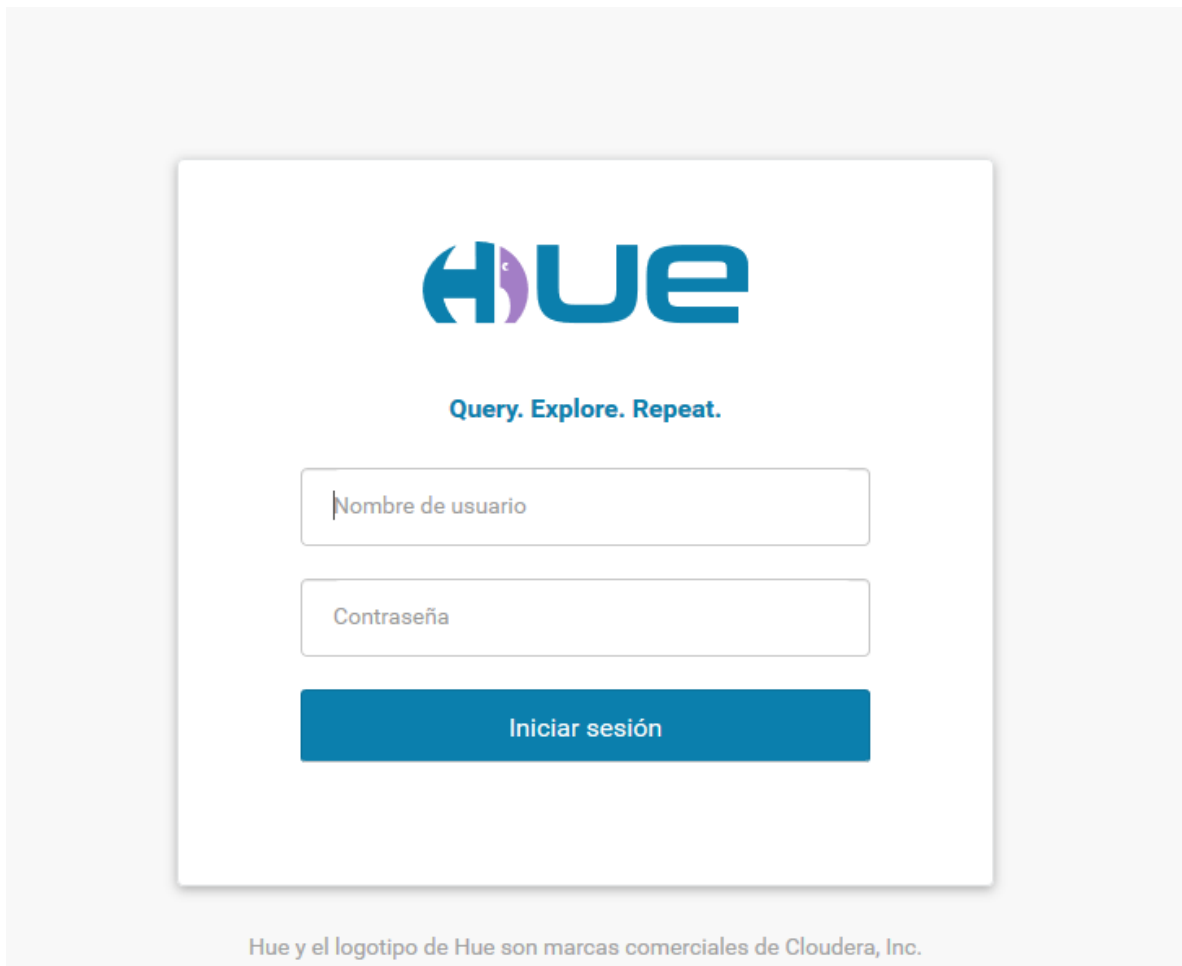
\$hdfs dfs -chmod 775 /bechmarks/archivohdfs.txt



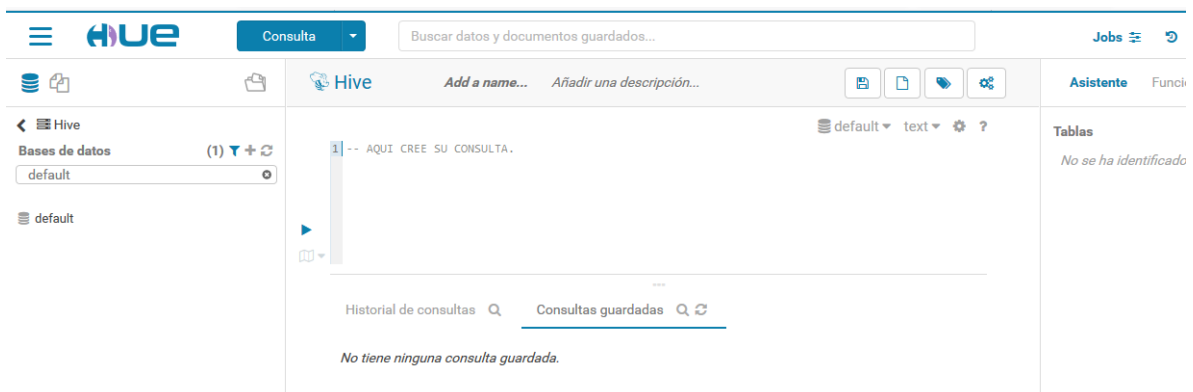
```
cloudera@quickstart:/home/cloudera/De
drwxr-xr-x - solr solr 0 2017-07-19 05:37 /solr
drwxrwxrwt - hdfs supergroup 0 2018-03-12 10:45 /tmp
drwxr-xr-x - hdfs supergroup 0 2018-06-21 16:36 /user
drwxr-xr-x - hdfs supergroup 0 2017-07-19 05:36 /var
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -ls /benchmarks/
Found 3 items
----- 1 root supergroup 1153 2018-06-24 13:45 /benchmarks/archiv
ohdfs.txt
drwxr-xr-x - root supergroup 0 2018-06-24 14:19 /benchmarks/master
drwxr-xr-x - root supergroup 0 2018-06-24 14:20 /benchmarks/stagin
g
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -chmod 775 /benchmarks/archi
vohdfs.txt
[root@quickstart CursoAvanzadoBigData]# hdfs dfs -ls /benchmarks/
Found 3 items
-rwxrwxr-x 1 root supergroup 1153 2018-06-24 13:45 /benchmarks/archiv
ohdfs.txt
drwxr-xr-x - root supergroup 0 2018-06-24 14:19 /benchmarks/master
drwxr-xr-x - root supergroup 0 2018-06-24 14:20 /benchmarks/stagin
g
[root@quickstart CursoAvanzadoBigData]#
```

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**:



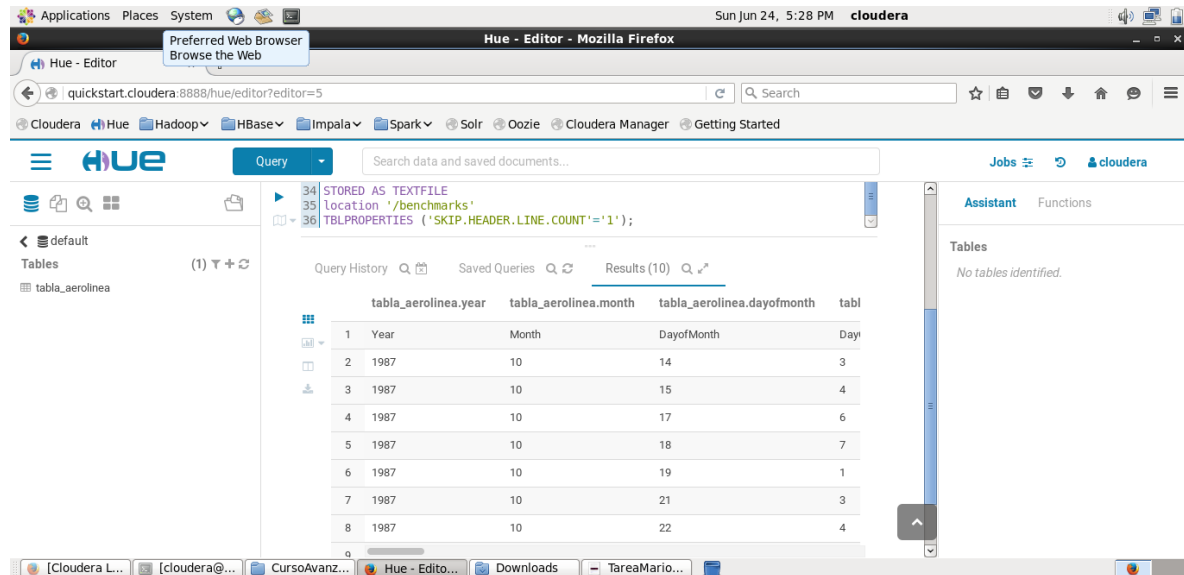
Entonces tome el siguiente código y cree una tabla en Hive:

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

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.

\$sudo service hue start

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.



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,

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

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.

Applications Places System Sun Jun 24, 9:36 PM cloudera

Change desktop appearance and behavior, get help, or log out Hue - Editor - Mozilla Firefox

about:sessionrestore x Hue - Editor x

quickstart.cloudera:8888/hue/editor?editor=8 Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Hue Query Search data and saved documents...

Jobs cloudera

Assistant Functions

Tables

No tables identified.

Query History Saved Queries Results (10)

	tabla_aerolinea.year	tabla_aerolinea.month	tabla_aerolinea.dayofmonth	tbl
1	NULL	NULL	NULL	NUL
2	1987	10	14	3
3	1987	10	15	4
4	1987	10	17	6
5				

```
CREATE EXTERNAL TABLE tabla_aerolinea(
```

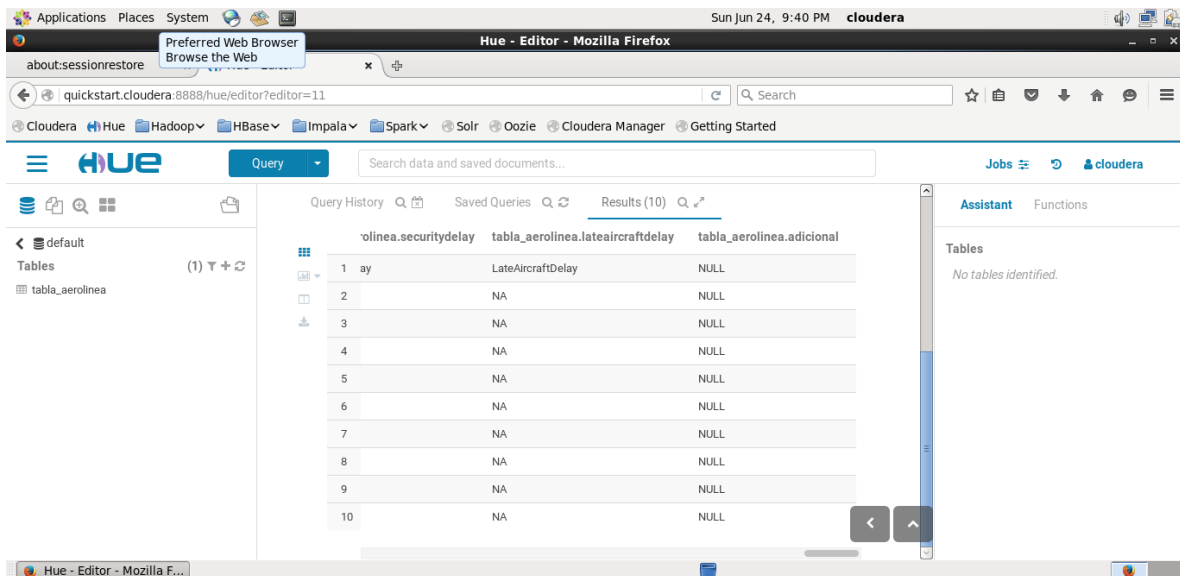
```
Year INT,  
Month INT,  
DayofMonth INT,  
DayOfWeek INT,  
DepTime INT,  
CRSDepTime INT,  
ArrTime INT,  
CRSArrTime INT,  
UniqueCarrier INT,  
FlightNum INT,  
TailNum INT,  
ActualElapsedTime INT,  
CRSElapsedTime INT,  
AirTime INT,  
ArrDelay INT,  
DepDelay INT,  
Origin INT,  
Dest INT,  
Distance INT,  
TaxiIn INT,  
TaxiOut INT,  
Cancelled INT,  
CancellationCode INT,  
Diverted INT,  
CarrierDelay INT,  
WeatherDelay INT,  
NASDelay INT,  
SecurityDelay INT,  
LateAircraftDelay INT)
```

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

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

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.

DROP TABLE tabla_aerolinea



The screenshot shows the Hue web interface in a Mozilla Firefox browser. The interface includes a top navigation bar with various tools like Cloudera, Hue, Hadoop, HBase, Impala, Spark, Solr, Oozie, Cloudera Manager, and Getting Started. The main area displays a query result table with 10 rows and 3 columns. The columns are labeled 'olinea.securitydelay', 'tabla_aerolinea.lateaircraftdelay', and 'tabla_aerolinea.adicional'. The first column contains values from 1 to 10, the second column contains 'LateAircraftDelay' for the first row and 'NA' for the others, and the third column contains 'NULL' for all rows. A sidebar on the left shows a tree view of tables, including 'default' and 'tabla_aerolinea'. A right sidebar shows an 'Assistant' panel with 'Functions' and 'Tables' sections, indicating 'No tables identified.'

olinea.securitydelay	tabla_aerolinea.lateaircraftdelay	tabla_aerolinea.adicional
1 ay	LateAircraftDelay	NULL
2	NA	NULL
3	NA	NULL
4	NA	NULL
5	NA	NULL
6	NA	NULL
7	NA	NULL
8	NA	NULL
9	NA	NULL
10	NA	NULL

La columna de tablas adicional se pone en nulo

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,

The screenshot displays the Hue web interface for running queries against a Hadoop cluster. At the top, there's a navigation bar with links like 'Applications', 'Places', 'System', and 'cloudera'. Below it, the browser address bar shows 'quickstart.cloudera:8888/hue/editor?editor=16'. The main interface has a sidebar on the left with icons for file operations and a 'Query' button. The central area contains a SQL query editor with the following code:

```

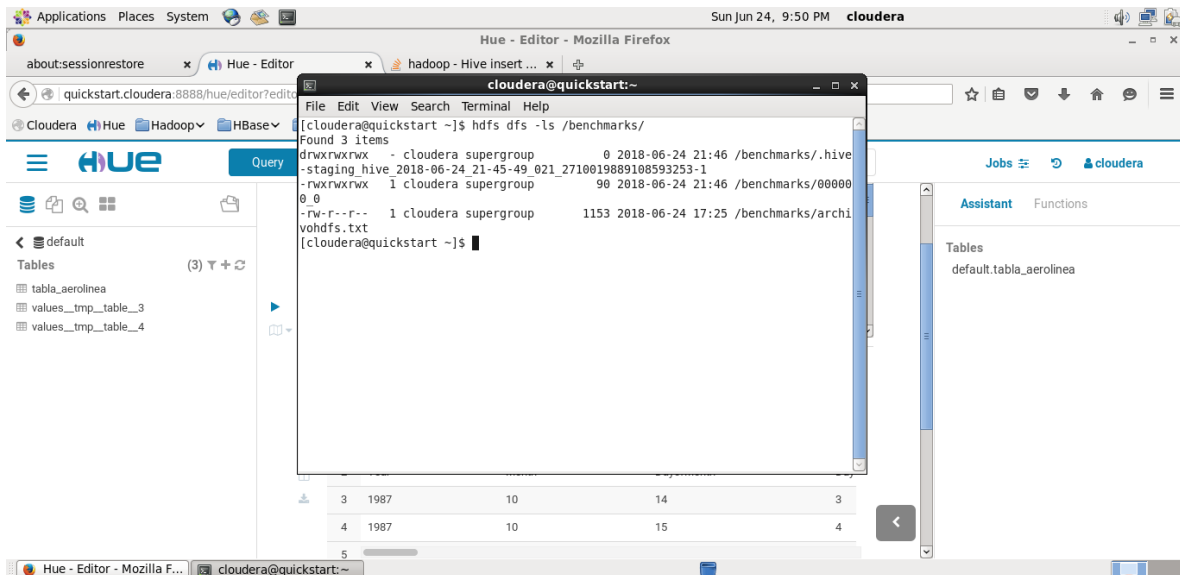
2 insert into table tabla_aerolinea values
3 ('NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA')
4 ('NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA')
5 ('NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA','NA')
6 select * from tabla_aerolinea
7 CREATE EXTERNAL TABLE tabla_aerolinea(
8   Year STRING,
9   Month STRING,
10  DayOfMonth STRING,
11  DayOfWeek STRING,

```

Below the query editor, there's a section for 'Query History', 'Saved Queries', and 'Results (11)'. The 'Results' tab is active, displaying a table with the following data:

	tabla_aerolinea.year	tabla_aerolinea.month	tabla_aerolinea.dayofmonth	tabla_aerolinea.day
1	NA	NA	NA	NA
2	Year	Month	DayOfMonth	Day
3	1987	10	14	3
4	1987	10	15	4
5				

On the right side of the interface, there are tabs for 'Assistant' and 'Functions', and a section titled 'Tables' listing 'default.tabla_aerolinea'.



Al parecer se almacena en un archivo temporal en el hdfs después de la inserción de los datos desde hue en cloudera.

SECCIÓN 3. PREGUNTAS ABIERTAS

20.- ¿Qué es el Sticky Bit? Ejemplifíquelo con el archivo **ejercicio_5.txt** adjuntando una captura de pantalla.

Este bit suele asignarse en directorios a los que todos los usuarios tienen acceso, y permite evitar que un usuario pueda borrar ficheros/directorios de otro usuario dentro de ese directorio, ya que todos tienen permiso de escritura. Seguro que lo estáis pensando, este bit se asigna siempre en /tmp y /var/tmp.

tmp tiene permisos 777, el bit sticky se asignaría del siguiente modo:

```
# chmod 1777 /tmp
```

También así:

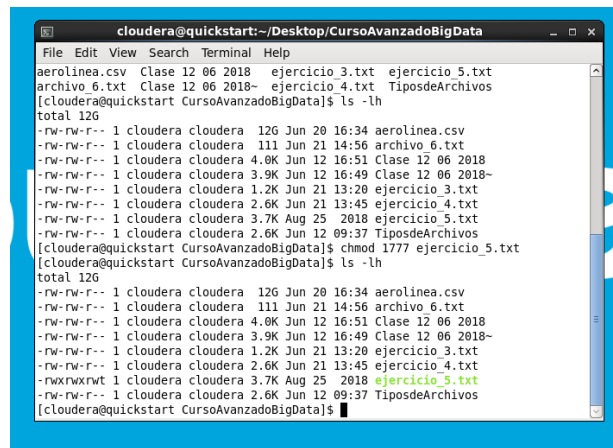
```
chmod o+t /tmp
```

Y para quitarlo:

```
chmod o-t /tmp
```

Si hacemos un ls veremos la "t" asignada:

```
drwxrwxrwt 13 root root 4096 2011-04-24 20:55 tmp
```



The screenshot shows a terminal window titled 'cloudera@quickstart: ~/Desktop/CursoAvanzadoBigData'. The user runs 'ls -lh' and then 'chmod 1777 ejercicio_5.txt'. The output shows the file 'ejercicio_5.txt' with permissions '-rwxrwxrwt' (highlighted in green), indicating that the 'sticky bit' is set. The terminal also shows a list of other files in the directory, including 'aerolinea.csv', 'archivo.6.txt', 'Clase 12 06 2018', 'ejercicio 3.txt', 'ejercicio 4.txt', and 'TiposdeArchivos'.

Fuente: <http://rm-rf.es/permisos-especiales-setuid-setgid-sticky-bit/>

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

NoSQL abarca una amplia gama de tecnologías y arquitecturas, busca resolver los problemas de escalabilidad y rendimiento de big data que las bases de datos relacionales no fueron diseñadas para abordar. NoSQL es especialmente útil cuando una empresa necesita acceder y analizar grandes cantidades de datos no estructurados o datos que se almacenan de forma remota en varios servidores virtuales en la nube. Contrariamente a las ideas falsas causadas por su nombre, NoSQL no prohíbe el lenguaje estructurado de consultas (SQL). Si bien es cierto que algunos sistemas NoSQL son totalmente no-relacionales, otros simplemente evitan funcionalidades relacionales seleccionadas como esquemas de tablas fijas y operaciones conjuntas. Por ejemplo, en lugar de utilizar tablas, una base de datos NoSQL podría organizar los datos en objetos, pares clave/valor o tuplas.

Podría decirse que la base de datos más popular NoSQL es Apache Cassandra. Cassandra, que una vez fue la base de datos propietaria de Facebook, fue liberada como código abierto en 2008. Otras implementaciones NoSQL incluyen SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemcacheDB y Voldemort. Las empresas que utilizan NoSQL incluyen Netflix, LinkedIn y Twitter. NoSQL se menciona a menudo en combinación con otras herramientas de big data, como el procesamiento paralelo masivo, las bases de datos a base de columnas y las bases de datos como servicio (DaaS). Hive e impala se utilizan para el tratamiento de grandes cantidades de información.

Fuente: <https://www.dezyre.com/questions/4863/what-is-the-difference-between-hive-and-hbase-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 comando nohup permite mantener la ejecución de un comando (el cual le pasamos como un argumento) pese a salir de la terminal (logout), ya que hace que se ejecute de forma independiente a la sesión. Básicamente, lo que hace es ignorar la señal HUP (señal que se envía a un proceso cuando la terminal que lo controla se cierra), esto implica que aunque cerremos la terminal, el proceso se siga ejecutando. La propia ayuda disponible en la shell (y en las páginas man) nos ayudará a entender el modo de ejecución del comando:

```
$ nohup --help
Modo de empleo: nohup ORDEN [ARGUMENTO]...
o bien: nohup OPCIÓN
Ejecuta ORDEN, descartando las señales de colgar.

--help      muestra esta ayuda y finaliza
--version   informa de la versión y finaliza

si la entrada estándar es una terminal, redirigirla desde /dev/null.
si la salida estándar es una terminal, añadir la salida a `nohup.out` si es posible,
en caso contrario a `${HOME}/nohup.out`.
si los errores van a una terminal, redirigirlos a la salida estándar.
Para guardar la salida a FILE, use `nohup COMMAND > FILE`
```

Un ejemplo sencillo sería la ejecución en segundo plano de un script cualquiera, gracias al comando nohup permitiremos la continuidad de la ejecución en caso de cualquier problema con la sesión, shell de ejecución, etc:

```
$ nohup ./miscript.sh &
```

Fuente : <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.	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.

Memoria Total : 4G

Memoria Utilizada: 2.7G

Memoria Libre = 1.2G

Total: Muestra la memoria RAM que tiene nuestro ordenador.

Used: Muestra el consumo de memoria RAM que están consumiendo los programas y procesos que se están ejecutando en nuestro ordenador.

Free: Muestra la memoria RAM que no estamos usando y por lo tanto está completamente libre sin realizar ninguna función. En otras palabras podemos decir que se trata de memoria RAM libre o “desperdiciada”.

Shared: Muestra la cantidad de memoria RAM que está siendo compartida y usada por más de un proceso o programa. De esta forma los procesos se pueden comunicar entre ellos y se evita copiar datos redundantes en la memoria.

buff/cache: Es la cantidad de memoria RAM que Linux se reserva para acelerar las lecturas en disco y para acelerar la asignación de memoria RAM a los programas.

Available: Es una estimación de la memoria RAM disponible para iniciar nuevos programas y procesos sin considerar la memoria Swap.

Fuente: <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?

El proyecto Apache [™] Hadoop[®] desarrolla software de código abierto para una computación distribuida confiable y escalable. La biblioteca de software Apache Hadoop es un marco que permite el procesamiento distribuido de grandes conjuntos de datos en clústeres de computadoras que usan modelos de programación simples. Está diseñado para escalar desde servidores únicos a miles de máquinas, cada una de las cuales ofrece cómputo y almacenamiento local. En lugar de confiar en el hardware para ofrecer alta disponibilidad, la biblioteca está diseñada para detectar y manejar fallas en la capa de aplicaciones, por lo que entrega un servicio altamente disponible sobre un grupo de computadoras, cada una de las cuales puede ser propensa a fallas.

Cloudera es el líder en software y servicios basados en Apache Hadoop y ofrece una nueva y poderosa plataforma de datos que permite a las empresas y organizaciones ver todos sus datos, estructurados y no estructurados, y hacer preguntas más grandes para una visión sin precedentes a la velocidad del pensamiento.

Fuente: <https://stackoverflow.com/questions/20139636/what-is-the-diff-between-apache-hadoop-and-cloudera-hadoop>

Cloudera es una empresa de nueva creación. Brindan soporte comercial para hadoop.

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

Sistema operativo	Tipos de sistemas de archivos admitidos
Dos	FAT16
Windows 95	FAT16
Windows95 OSR2	FAT16, FAT32
Windows 98	FAT16, FAT32
Windows NT4	FAT, NTFS (versión 4)
Windows 2000/XP	FAT, FAT16, FAT32, NTFS (versiones 4 y 5)
Linux	Ext2, Ext3, ReiserFS, Linux Swap (FAT16, FAT32, NTFS)
MacOS	HFS (Sistema de Archivos Jerárquico), MFS (Sistemas de Archivos Macintosh)
OS/2	HPFS (Sistema de Archivos de Alto Rendimiento)
SGI IRIX	XFS
FreeBSD, OpenBSD	UFS (Sistema de Archivos Unix)
Sun Solaris	UFS (Sistema de Archivos Unix)
IBM AIX	JFS (Sistema Diario de Archivos)

Fuente: <http://oskarj023.blogspot.com/2012/>

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

Serde es una interfaz de hive una interfaz que permite indicarle a Hive como debe procesar un registro. SerDe es una combinación de Serializer y Deserializer.

- Deserializer toma una representación string o binaria y lo convierte a un objeto Java que Hive puede manipular.

- Serializer: toma un objeto Java y lo convierte en algo que Hive puede escribir a HDFS. Para usar un SerDe a la hora de crear la tabla debo indicar que SerDe usar:

```

ADD JAR /tmp/hive-serdes-1.0-SNAPSHOT.jar

CREATE EXTERNAL TABLE tweets (
  ...
  retweeted_status STRUCT<
    text:STRING,
    user:STRUCT<screen_name:STRING,name:STRING>>,
  entities STRUCT<
    urls:ARRAY<STRUCT<expanded_url:STRING>>,
    user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
    hashtags:ARRAY<STRUCT<text:STRING>>>,
    text STRING,
  ...
)
PARTITIONED BY (datehour INT)
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/flume/tweets';

```

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

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

BigTable es un sistema de gestión de base de datos creado por Google con las características de ser: distribuido, de alta eficiencia y propietario. Está construido sobre GFS (Google File System), Chubby Lock Service, y algunos otros servicios y programas de Google, y funciona sobre 'commodity hardware' (sencillos y baratos PCs con procesadores Intel). BigTable comenzó a ser desarrollado a principios de 2004. BigTable almacena la información en tablas multidimensionales cuyas celdas están, en su mayoría, sin utilizar. Además, estas celdas disponen de versiones temporales de sus valores, con lo que se puede hacer un seguimiento de los valores que han tomado históricamente. Para poder manejar la información, las tablas se dividen por columnas, y son almacenadas como 'tabletas' de unos 200 Mbytes cada una. Cada máquina almacena 100 tabletas, mediante el sistema 'Google File System'. La disposición permite un sistema de balanceo de carga (si una tableta está recibiendo un montón de peticiones, la máquina puede desprenderse del resto de las tabletas o trasladar la tableta en cuestión a otra máquina) y una rápida recomposición del sistema si una máquina 'se cae'.

BigQuery es un servicio web RESTful que permite el análisis interactivo de grandes conjuntos de datos que trabajan en conjunto con Google Storage. Es una Infraestructura como Servicio (IaaS) que puede usarse de forma complementaria con MapReduce.

Fuente:

<https://www.google.com.mx/search?q=traductor+google&oq=traductor+google&aqs=chrome.69l67j0j9&sourceid=chrome&ie=UTF-8>

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

Un Data Lake es un repositorio donde se almacenan todos los datos de la compañía, una gran cantidad de datos en bruto, estructurados y sin estructurar, sin ningún tipo de pre procesamiento (raw data) y sin ningún tipo de esquema, se mantienen allí almacenados los que son necesarios para ser analizados.

La información que se almacena en el Data Lake procede de diversas fuentes de datos, por lo que guarda datos de todo tipo: procedentes de bases de datos, documentos ofimáticos, registros de servidores, recursos extraídos de Internet, redes sociales, textos, etc. con el objetivo de ser estudiados y analizados posteriormente.

Las empresas vierten los datos en estos almacenes y los recuperan cuando son necesarios. Es en ese momento, cuando las empresas tienen la necesidad de los datos, que estos son ordenados y se diseña una estructura de análisis apropiada. Podríamos describir el data lake como un almacenamiento de bajo coste y el acceso a la información original es directo al disponer de todos los datos en bruto.

Un data lake funciona de la siguiente manera: se asigna un identificador único a cada elemento del data lake y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando en la empresa se presenta una cuestión sobre el negocio que debemos resolver y requerimos de datos, podemos solicitarle al data lake los datos que están relacionados con esa cuestión. Una vez obtenidos podemos analizar ese conjunto de datos más pequeño para ayudar a obtener una respuesta.

Data Lake vs Big Data

Es posible que hayas relacionado el término Data Lake con Big Data, aunque ambos son almacenes de datos operan de formas diferentes.

En Big Data se recoge información procedente de diversas fuentes también, pero que se filtra se organiza y almacena para ser analizada de inmediato con un objetivo concreto. Se trata de un formato estructurado que trabaja a corto plazo y tiene en cuenta solo lo que es útil en el momento, el resto de datos que no son necesarios para el análisis que se lleva a cabo en ese momento se desechan. Analiza los datos solamente una vez según la estructura de análisis fijada y exporta los resultados válidos.

En Data Lake se recoge la información y se almacena, pero, a diferencia de Big Data, no se ordena, ni se filtra, ni se organiza, en el momento de almacenar la información no se produce ninguna alteración respecto a la información original. La información almacenada será analizada cuando se necesite. Por este motivo la información almacenada será útil siempre que se necesite, independientemente de que cambie el objetivo u orientación del análisis, y los datos se podrán volver a analizar tantas veces como se requieran. Sin embargo, a diferencia de Big Data, un data lake requiere de mucho más espacio de almacenamiento porque su cantidad de datos es indefinida y va creciendo.

Cada vez más, el término data lake está siendo aceptado como una forma de describir cualquier gran conjunto de datos en el que el esquema y los requisitos de datos no se definen hasta que los datos se consultan.

La arquitectura utilizada para el almacenaje de los datos en data lake es una arquitectura plana.

Un data warehouse es un repositorio unificado para todos los datos que recogen los diversos sistemas de una empresa. El repositorio puede ser físico o lógico y hace hincapié en la captura de datos de diversas fuentes sobre todo para fines analíticos y de acceso.

Normalmente, un data warehouse se aloja en un servidor corporativo o cada vez más, en la nube. Los datos de diferentes aplicaciones de procesamiento de transacciones Online (OLTP) y otras fuentes se extraen selectivamente para su uso por aplicaciones analíticas y de consultas por usuarios.

Data Warehouse es una arquitectura de almacenamiento de datos que permite a los ejecutivos de negocios organizar, comprender y utilizar sus datos para tomar decisiones estratégicas. Un data warehouse es una arquitectura conocida ya en muchas empresas modernas.

Estructuras de un Data Warehouse

La arquitectura de un data warehouse puede ser dividida en tres estructuras simplificadas: básica, básica con un área de ensayo y básica con área de ensayo y data marts.

Con una estructura básica, sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con metadatos. Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.

Al añadir un área de ensayo que se puede colocar entre las fuentes de datos y el almacén, ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización.

Se puede hacer agregando data marts, que son sistemas diseñados para una línea de negocio en particular. Se pueden tener data marts separados para ventas, inventario y compras, por ejemplo, y los usuarios finales pueden acceder a datos de uno o de todos los data marts del departamento.

Fuente: <https://www.powerdata.es/data-warehouse>

<https://www.deustoformacion.com/blog/gestion-empresas/que-es-para-que-sirve-data-lake>

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

NFS

Primer sistema comercial de archivos en red ([Sun Microsystems](#), 1984) estándar, multiplataforma que permite acceder y compartir archivos en una red [C/S](#) heterogénea como si estuvieran en un solo disco, [es decir](#), montar un directorio de una máquina remota en una máquina local.

AFS:

El Andrew file system es un sistema de archivos distribuido comercial (CMU 1983, Transarc 1989, IBM 1998) para compartir archivos de manera transparente, escalable e independiente de la ubicación real.

Implementaciones de AFS:

- [OpenAFS](#): Versión open-source de AFS (IBM 2000).
- XCoda: Sistema de archivos distribuido experimental open-source derivativo de AFS (CMU 1987). Se distingue por soportar dispositivos móviles.

DCE DFS

DCE Distributed File System es un [sistema de ficheros](#) distribuido de [DCE](#) que permite agrupar archivos repartidos en diferentes máquinas, en un espacio de nombres único. Está basado casi por completo en el sistema de ficheros [AFS](#) pero con ligeras diferencias.

Fuente: https://es.wikipedia.org/wiki/Sistema_de_archivos_distribuido

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:

```
Yum install -y  
http://dl.fedoraproject.org/pub/epel/7/x86_64/Packages/e/epel-release-7-11.noarch.rpm
```

```
yum install -y python-pip python-devel python-virtualenv  
yum groupinstall 'Development Tools'  
virtualenv jupyter-virtualenv  
source jupyter-virtualenv/bin/activate  
pip install jupyter
```



Logout

FilesRunningClusters

Select items to perform actions on them.

UploadNew

Downloads

NameLast Modified

...

64bit-master

_MACOSX

aaronMongo

apache-maven-3.5.3-bin

BBVAWorkbench

cloudera-quickstart-vm-5.12.0-0-virtualbox

codigo_completo

Compilación Calculadora

Curso Avanzado

Datio

DB_VIS_201701

Downloads

seconds ago

2 months ago

a year ago

3 months ago

2 months ago

5 months ago

5 months ago

2 months ago

5 days ago

a day ago

2 months ago

6 months ago

3 months ago

Instalación de Jupyter