Ejercicios Hive

1. Entrar en Hive
   "hive"
2. Modificar la propiedad correspondiente para mostrar por pantalla las cabeceras de las tablas
   "set hive.cli.print.header=true;"
3. Crear una base de datos llamada "cursohivedb"
   drop database if exists userdb cascade;
   show databases; (mostrar las que tenemos)
   "CREATE DATABASE cursohivedb;"
4. Situarnos en la base de datos recién creada para trabajar con ella
   "USE cursohivedb"
5. Comprobar que la base de datos está vacía
   "SHOW TABLES"
6. Crear una tabla llamada "iris" en nuestra base de datos que contenga 5 columnas (s_length float,s_width float,p_length float,p_width float,clase string) cuyos campos estén separados por comas (ROW FORMAT DELIMITED FIELDS TERMINATED BY ',')

```
DROP TABLE iris;
create table iris(
s_length float,
s_width float,
p_length float,
p_width float,
clase string
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY
',';
```

7. Comprobar que la tabla se ha creado y el tipado de sus columnas
   "SHOW TABLES;"
   "DESC iris;"
8. Importar el fichero "iris_completo.txt" al local file system del cluster en la carpeta /home/cloudera/ejercicios/ejercicios_HIVE
9. Copiar el fichero a HDFS en la ruta /user/cloudera/hive. Reailzar las acciones necesarias
   "hadoop fs -mkdir /user/cloudera/hive"
   "hadoop fs -put /home/cloudera/ejercicios/ejercicios_HIVE/iris_completo.txt /user/cloudera/hive"
10. Comprueba que el fichero está en la ruta en HDFS indicada
11. Importa el fichero en la tabla iris que acabamos de crear desde HDFS
    "load data inpath '/user/cloudera/hive/iris_completo.txt' into table iris;"
12. Comprobar que la table tiene datos
13. Mostrar las 5 primeras filas de la tabla iris
    "Select * from iris limit 5;"

14. Mostrar solo aquellas filas cuyo s_length sea mayor que 5. Observad que se ejecuta un MapReduce y que el tiempo de ejecución es un poco mayor

```
"Select * from iris where s_length>5;"
```

15. Seleccionar la media de s_width agrupados por clase. Observad que ahora el tiempo de ejecución aumenta considerablemente.

```
"Select avg(s_width) from iris group by clase;"
```

16. Pregunta: vemos que aparece un valor NULL como resultado en la query anterior. ¿Por qué? ¿cómo los eliminarías?
17. Insertar en la tabla la siguiente fila (1.0,3.2,4.3,5.7,"Iris-virginica")

```
"insert into table iris values
(1.0,3.2,4.3,5.7,"Iris-virginica");"
```

18. Contar el número de ocurrencias de cada clase

```
"select clase, count(*)
from iris
group by clase"
```

19. Seleccionar las clases que tengan más de 45 ocurrencias

```
"select clase, count(*)
from iris
group by clase
having count(*) >45;"
```

20. Utilizando la función LEAD, ejecutar una query que devuelva la clase, p_length y el LEAD de p_length con Offset=1 y Default_Value =0, particionado por clase y ordenado por p_length.

```
select clase,
p_length,
LEAD(p_length,1,0) OVER (PARTITION BY clase ORDER BY
p_length) as Lead
from iris;
```

21. Utilizando funciones de ventanas, seleccionar la clase, p_length, s_length, p_width, el número de valores distintos de p_length en todo el dataset, el valor máximo de s_length por clase y la media de p_width por clase, ordenado por clase y s_length de manera descendente.

```
"select clase,
p_length,
s_length,
p_width,
count(p_length) over (partition by p_length) as
pl_ct,
max(s_length) over (partition by clase) as sl_ct,
avg(p_width) over (partition by clase) as sl_av
from iris
order by clase,s_length desc;"
```