Ejercicios Hive

* Entrar en Hive

hive

* Modificar la propiedad correspondiente para mostrar por pantalla las cabeceras de las tablas

set hive.cli.print.header=true;

* Crear una base de datos llamada “cursohivedb”

create database if not exists cursohivedb;

* Situarnos en la base de datos recién creada para trabajar con ella

use cursohivedb;

* Comprobar que la base de datos está vacía

show tables;

* 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 ',')

create table if not exists cursohivedb.iris(

s\_length float, s\_width float,

p\_length float, p\_width float,

clase string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

* Comprobar que la tabla se ha creado y el tipado de sus columnas

show tables;

desc iris;

* Importar el fichero “iris\_completo.txt” al local file system del cluster en la carpeta /home/cloudera/ejercicios/ejercicios\_HIVE
* Copiar el fichero a HDFS en la ruta /user/cloudera/hive. Realizar las acciones necesarias

exit;

hadoop fs -mkdir /user/cloudera/hive;

hadoop fs -put /home/cloudera/ejercicios/iris\_completo.txt /user/cloudera/hive;

* Comprueba que el fichero está en la ruta en HDFS indicada

hadoop fs -ls /user/cloudera/hive

* Importa el fichero en la tabla iris que acabamos de crear desde HDFS

hive

load data inpath '/user/cloudera/hive/iris\_completo.txt' into table cursohivedb.iris;

* Comprobar que la table tiene datos

select \* from cursohivedb.iris;

* Mostrar las 5 primeras filas de la tabla iris

select \* from cursohivedb.iris limit 5;

* 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 cursohivedb.iris as iris where iris.s\_length>5;

* Seleccionar la media de s\_width agrupados por clase. Observad que ahora el tiempo de ejecución aumenta considerablemente.

select avg(s\_width) from cursohivedb.iris GROUP BY clase;

* Pregunta: vemos que aparece un valor NULL como resultado en la query anterior. ¿Por qué? ¿cómo los eliminarías?

Porque algún dato de la columna s\_width, es un valor nulo o incompatible, para ello podemos ignorar nulos usando la siguiente línea:

select avg(s\_width) from cursohivedb.iris WHERE s\_width is not null GROUP BY clase;

* Insertar en la tabla la siguiente fila (1.0,3.2,4.3,5.7,"Iris-virginica")

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

* Contar el número de ocurrencias de cada clase

select count(clase) from cursohivedb.iris GROUP BY clase;

* Seleccionar las clases que tengan más de 45 ocurrencias

select count(clase) from cursohivedb.iris GROUP BY clase;

* 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 cursohivedb.iris;

* 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 conteo\_p, max(s\_length) over (PARTITION BY clase) as maxs\_length, avg(p\_width) over (PARTITION BY clase) as mediapwidth from cursohivedb.iris ORDER BY clase, s\_length DESC;