Ejercicios Hive

Realizados por Paula Iglesias en la semana 2 de la formación - Dia 10/6/2021

Resumen comandos más utilizados:

Poner el formato del teclado en esp (en máquina debian)

setxkbmap -layout 'es,es' -model pc105

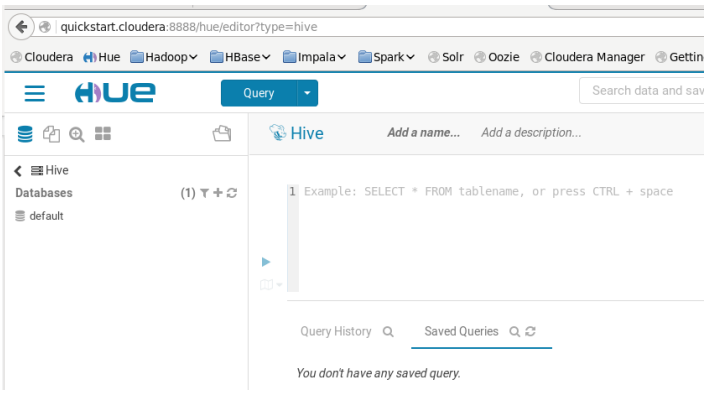
Crear carpeta "ejercicios" en Cloudera

hadoop fs -mkdir /user/cloudera/ejercicios

Ejercicios Hive

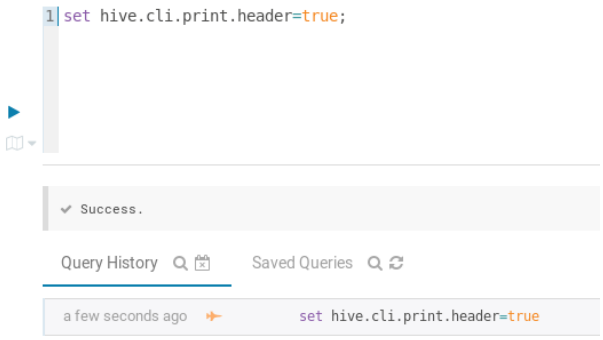
1. Entrar en Hive
    a. `hive`

    He entrado a la interfaz de Hive en la MV Cloudera:



2. Modificar la propiedad correspondiente para mostrar por pantalla las cabeceras de las tablas
    a. `"set hive.cli.print.header=true;"`

    Succcess:



3. ‘,Crear una base de datos llamada “cursohivedb”

a. `Créate database cursohivedb`

4. Situarnos en la base de datos recién creada para trabajar con ella
   a. `Use cursohivedb`
5. Comprobar que la base de datos está vacía
   a. `Show tables;`

   `Resultado: 0 results`

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 ',')

```
 CREATE TABLE iris (
S_length float,
S_width float,
P_length float,
P_width float,
Clase string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY
'id'
```

CREATE TABLE iris (

S_length float,

S_width float,

P_length float,

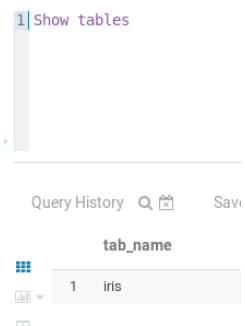
P_width float,

Clase string

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

7. Comprobar que la tabla se ha creado y el tipado de sus columnas
   a. `Show tables`



   b. `Desc iris`

```
1 Desc iris
```

Query History | Saved Queries | Results (5)

	col_name	data_type
1	s_length	float
2	s_width	float
3	p_length	float
4	p_width	float
5	clase	string

8. Importar el fichero “iris_completo.txt” al local file system del cluster en la carpeta /home/cloudera/ejercicios/ejercicios_HIVE

hadoop fs -put /mnt/Shared/iris_completo.txt /user/cloudera/hive

   a. Copiar el fichero a HDFS en la ruta /user/cloudera/hive. Realizar las acciones necesarias

```
Hadoop fs -mkdir /user/cloudera/ejercicios/ejercicios_hive
```

   b. `Hadoop fs -put /home/Cloudera/ejercicios/ejercicios_hive/iris_completo.txt /user/Cloudera/hive`

9. Comprueba que el fichero está en la ruta en HDFS indicada

```
[root@quickstart cloudera]# hadoop fs -ls  /user/cloudera/hive
Found 1 items
-rw-r--r--   1 root cloudera       4551 2021-06-10 05:21 /user/cloudera/hive/iri
s completo.txt
```

10. Importa el fichero en la tabla iris que acabamos de crear desde HDFS
    a. `Load data inpath '/user/cloudera/hive/iris_completo.txt' into table iris; // Desde Hive`
11. Comprobar que la table tiene datos

```
1 select * from
2 iris;
```

Query History | Saved Queries | Results (100+)

	iris.s_length	iris.s_width	iris.p_length	iris.p_width	iris.clase
1	5.0999999046325684	3.5	1.3999999761581421	0.20000000298023224	Iris-setosa
2	4.9000000953674316	3	1.3999999761581421	0.20000000298023224	Iris-setosa
3	4.6999998092651367	3.2000000476837158	1.2999999523162842	0.20000000298023224	Iris-setosa
4	4.5999999046325684	3.0999999046325684	1.5	0.20000000298023224	Iris-setosa
5	5	3.5999999046325684	1.3999999761581421	0.20000000298023224	Iris-setosa
6	5.4000000953674316	3.9000000953674316	1.7000000476837158	0.40000000596046448	Iris-setosa
7	4.5999999046325684	3.4000000953674316	1.3999999761581421	0.30000001192092896	Iris-setosa
8	5	3.4000000953674316	1.5	0.20000000298023224	Iris-setosa
9	4.4000000953674316	2.9000000953674316	1.3999999761581421	0.20000000298023224	Iris-setosa
10	4.9000000953674316	3.0999999046325684	1.5	0.10000000149011612	Iris-setosa
11	5.4000000953674316	3.7000000476837158	1.5	0.20000000298023224	Iris-setosa

12. Mostrar las 5 primeras filas de la tabla iris
    a. `Select * from iris limit 5;`

13. 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
    a. `Select * from iris as i where i.s_length>5`
14. Seleccionar la media de s_width agrupados por clase. Observad que ahora el tiempo de ejecución aumenta considerablemente.
    a. `Select avg(s_width) from iris GROUP BY clase;`
15. Pregunta: vemos que aparece un valor NULL como resultado en la query anterior. ¿Por qué? ¿cómo los eliminarías?

    Porque había algún dato erróneo, no numérico o nulo en el campo de alguna clase. Para eliminarlos podríamos añadir la condición where para que fuera distinto de null.

16. Insertar en la tabla la siguiente fila (1.0,3.2,4.3,5.7,"Iris-virginica")
    a. `Insert into table iris values (`1.0,3.2,4.3,5.7,"Iris-virginica"`)`
17. Contar el número de ocurrencias de cada clase
    a. `Select count(clase) from iris group by clase;`
18. Seleccionar las clases que tengan más de 45 ocurrencias
    a.
    ```
    Select clase from iris group by clase having
    count(*)>45;
    ```
19. 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.
    a. select clase, p_length, LEAD(p_length,1,0) OVER (PARTITION BY clase ORDER BY p_length) as Lead from iris;

    `LEAD muestra la siguiente fila`

20. 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.

    a.
    ```
    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;
    ```