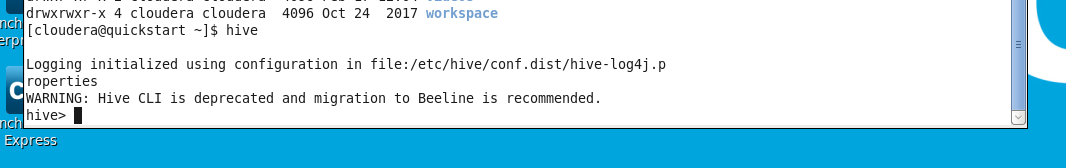
HIVE

Es una infraestructura para el almacenaje y consultade datos basada en Hadoop

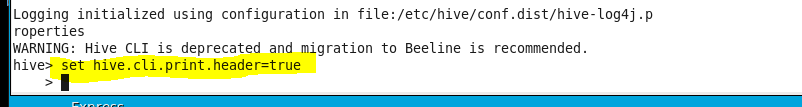
Hadoop proporciona escalabilidad masiva y con capacidades de tolerancia a fallos para el procesamiento y almacenamiento de datos

Ejercicios Hive

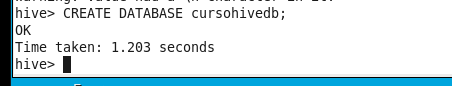
1. Entrar en Hive

2. Modificar la propiedad correspondiente para mostrar por pantalla las cabeceras de las

tablas



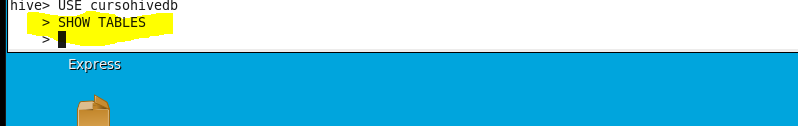
3. Crear una base de datos llamada “cursohivedb”



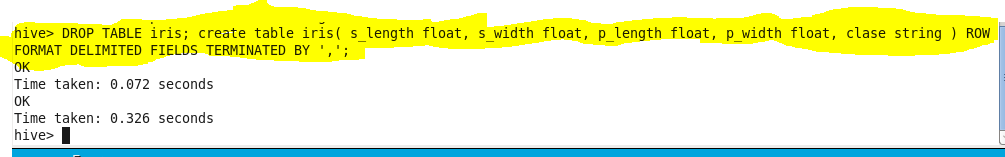
4. Situarnos en la base de datos recién creada para trabajar con ella



5. Comprobar que la base de datos está vacía

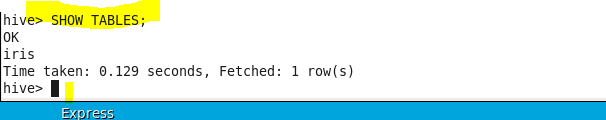


6. Crear una tabla llamada “iris” en nuestra base de datos que contenga 5 columnas

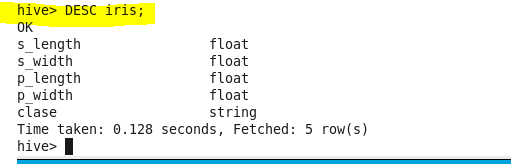


7. Comprobar que la tabla se ha

creado y el tipado de sus columnas



Tipado de sus columnas



8. Importar el fichero

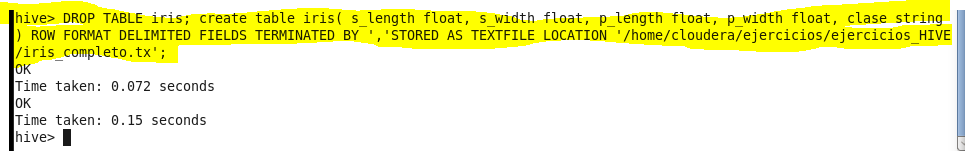
“iris\_completo.txt” al local file

system del cluster en la carpeta

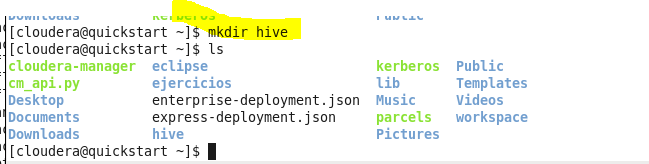
/home/cloudera/ejercicios/ejercicios\_HIVE

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

Necesarias

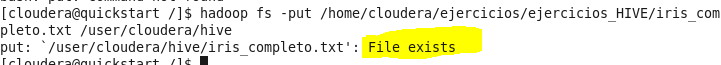


9. Comprueba que el fichero está en la ruta en HDFS indicada

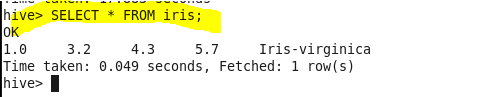




10. Importa el fichero en la tabla iris que acabamos de crear desde HDFS



11. Comprobar que la table tiene datos



12. Mostrar las 5 primeras filas de la tabla iris

Hive> 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

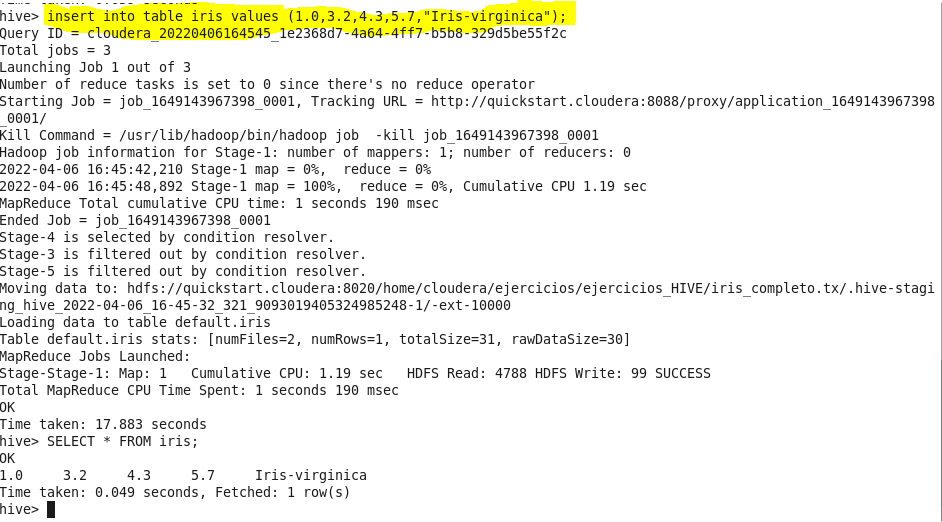
Hive> Select \* from iris as i where i.s\_length>5

14. Seleccionar la media de s\_width agrupados por clase. Observad que ahora el tiempo

de ejecución aumenta considerablemente.

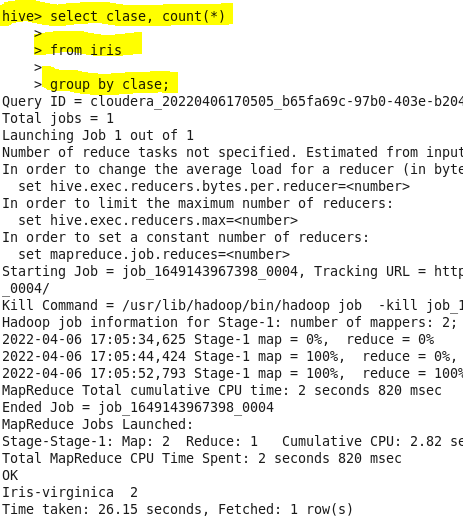
15. Pregunta: vemos que aparece un valor NULL como resultado en la query anterior. ¿Por qué? ¿cómo los eliminarías? Porque había algún dato erróneo, no numérico o nulo en el campo de alguna clase. Para eliminarlos podríamos añadir la condición where para que fuera distinto de null.

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

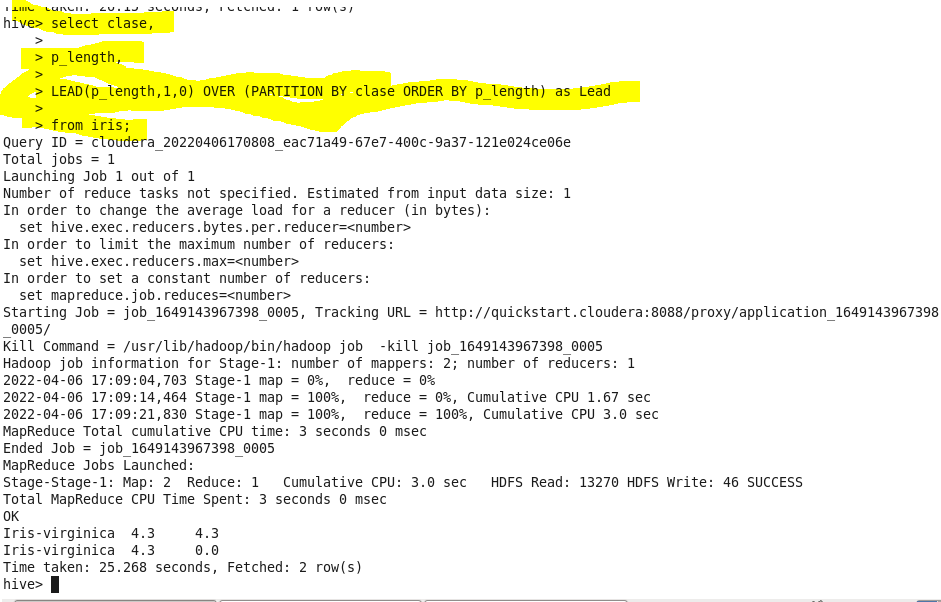


17. Contar el número de ocurrencias de cada clase

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.



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.



