EJERCICIOS DE HIVE

Estructuras de datos

Partiendo de la discografía de Pink Floyd (año, nombre disco, ranking EEUU, ranking UK)

1967, The Piper at the Gates of Dawn,131,6

1968, A Saucerful of Secrets,999,9

1969, Music from the Film More,153,9

1969, Ummagumma,74,5

1970, Atom Heart Mother,55,1

1972, Obscured by Clouds, 46,6

1973, The Dark Side of the Moon, 1,1

1975, Wish you Were Here, 1,1

1977, Animals, 3,2

1979, The Wall, 1,3

1983, The Final Cut, 6,1

1987, A Momentary Lapse of Reason,3,3

1994, The Division Bell, 1,1

2014, The Endless River, 3, 1

Indicar los comandos empleados para resolver las siguientes preguntas:

1. Crear un fichero de texto con la información anterior (IMPORTANTE: al crear el fichero tener cuidado con los caracteres al final de línea)

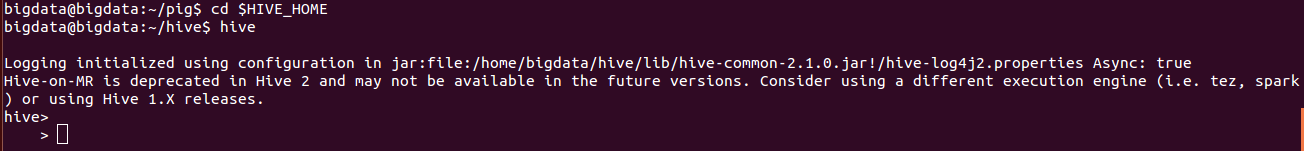
Creamos una carpeta llamada “ejercicioListaExitos” y dentro de ella creamos el fichero con los datos:



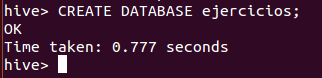
Nota: He reutilizado el fichero que había creado para el ejercicio anterior.

2. Acceder a Hive y crear una base de datos llamada ejercicios

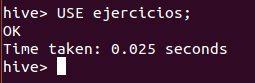
Entramos en Hive:



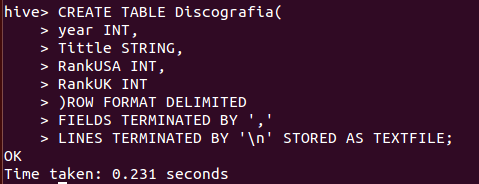
Y creamos la base de datos



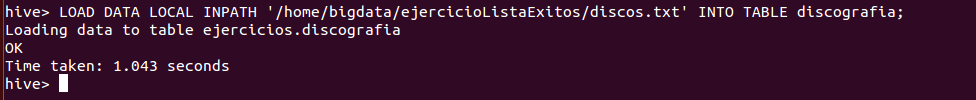
3. Usar la base de datos anterior



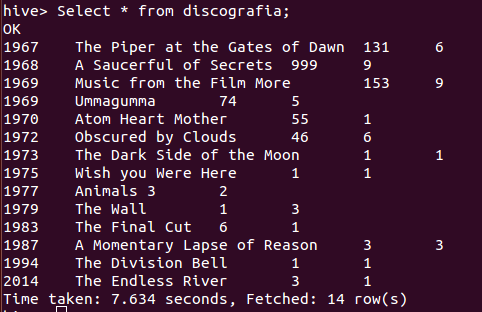
4. Crear una tabla en Hive en la base de datos anterior que permita almacenar los datos anteriores indicando que el formato de separación es como el siguiente de tipo tabulación (create table …… (...........) row format delimited fields terminated by ',' stored as textfile;)



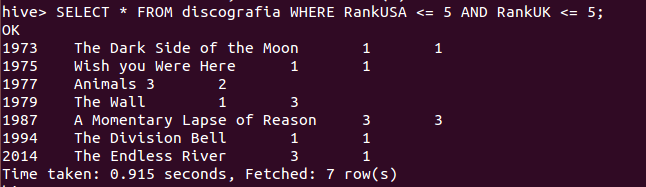
5. Cargar el fichero de texto



6. Acceder a Hive y ejecutar un consulta sencilla (select \*) para verificar que hay datos y se han cargado correctamente. En caso contrario, volver a cargar los datos

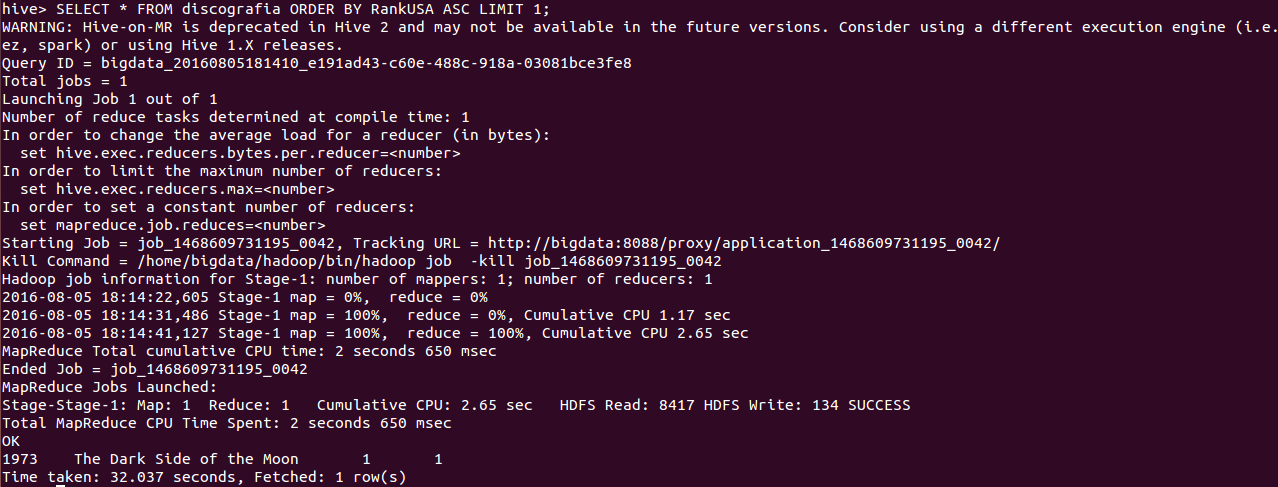


7. Calcular los discos que estuvieron a la vez entre los 5 primeros lugares en EEUU y UK



8. (OPCIONAL) Obtener la máxima y mínima posición que ocuparon los discos de Pink Floyd en EEUU y en UK (por ejemplo empleando el comando order y limit en dos sentencias)

Máxima posición en USA:



Mínima posición en USA:





Máxima posición en UK:





Mínima posición en UK:





9. (OPCIONAL) Repetir todos los ejercicios empleando una tabla con estructuras de datos complejas

UDFS

Acceder a la página

<http://www.minhap.gob.es/es-ES/Areas%20Tematicas/Administracion%20Electronica/OVEELL/Paginas/DeudaViva.aspx>

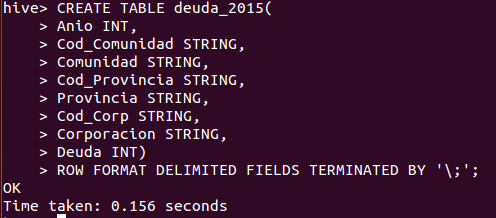
y descargar el fichero de deuda de ayuntamientos del año 2015.

Dscargamos el fichero y limpiamos la filas que no necesitamos. Lo guardamos como CSV utilizando “;” como separador.

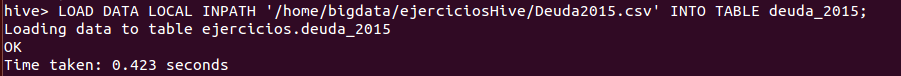
Una vez descargado y abierto para analizar su información, resolver los siguientes ejercicios:

1. Indicar el comando de creación de una tabla en Hive que contenga la información de la solapa datos con las siguientes columnas:

* Anio
* Cod\_Comunidad
* Comunidad
* Cod\_Provincia
* Provincia
* Cod\_corp
* Corporacion
* Deuda



2. Cargar el fichero anterior descargado en esa tabla e indicar el comando empleado



3. Obtener la corporación que tiene la tercera deuda más grande



Obtenemos las tres corporaciones con mayor deuda. La que buscamos es la tercera (Barcelona):



4. Obtener la corporación que tiene la deuda más grande de Andalucía





OPCIONALES

5. Obtener la suma de deudas de cada comunidad autónoma ordenadas por comunidad

6. Obtener la suma de las deudas de las corporaciones que empiezan por vocal

7. Exportar las corporaciones cuya deuda sea mayor a 10.000 y pertenezcan a Cataluña a un fichero llamado deudas y adjuntarlo. Puedes emplear un comando similar al siguiente:

INSERT OVERWRITE LOCAL DIRECTORY '/home/bigdata/ejemplosHive/RESULTADO/' row format delimited fields terminated by ',' select \* from .........