EJERCICIOS DE PIG

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

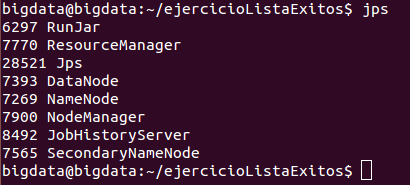
1. Crear un fichero llamado discos.txt

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



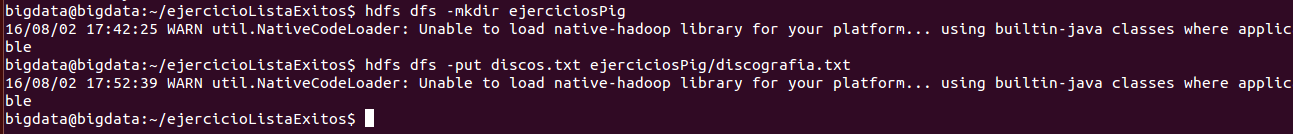
2. Arrancar HDFS, Yarn y el job history

Comprobamos que los demonios están arrancados:



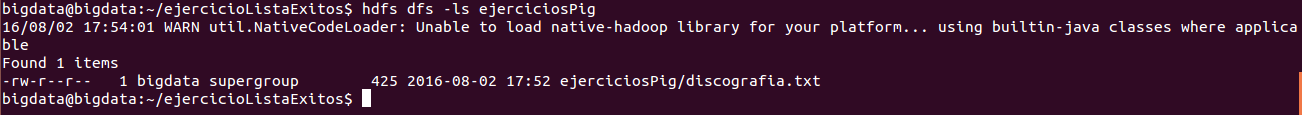
3. Subir el fichero a HDFS dentro de la carpeta /ejerciciosPig/discografia.txt

Creamos la carpeta en el sistema hdfs y subimos el archivo:

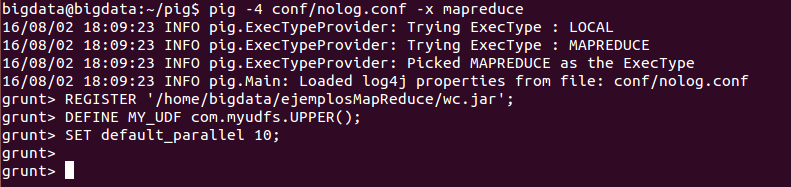


4. Ejecutar la instrucción ls sobre Hadoop para indicar el tamaño del fichero

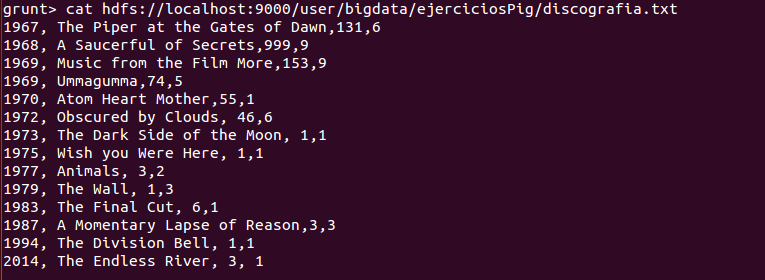
Ejecutamos un ls sobre la carpeta para comprobar que está subido y ver el tamaño:



5. Arrancar pig en modo distribuido (si se desea eliminar trazas de log) y ejecutar el siguiente comando: cat hdfs://localhost:9000/ejerciciosPig/discografia.txt para confirmar que los primeros puntos han funcionado correctamente y el fichero está subido a HDFS



El fichero no se encuentra en la ruta indicada sino dentro de la carpeta /user/bigdata:

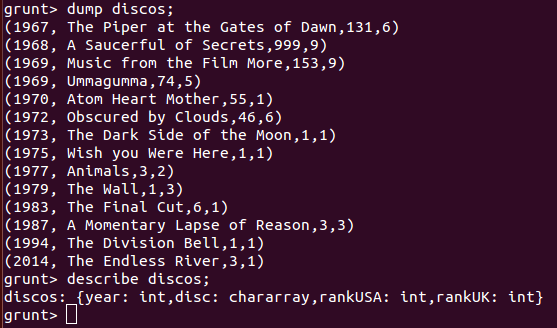


6. Cargar el fichero de hdfs en una variable llamada discos

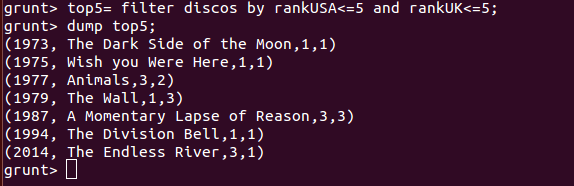
Cargamos el fichero en la variable:



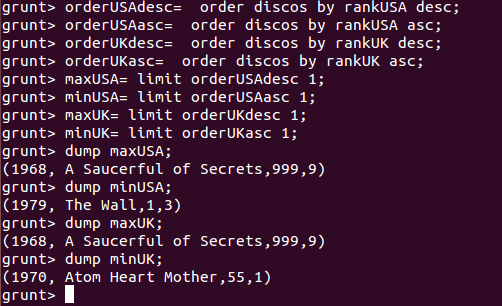
Y lo comprobamos:



7. Calcular los discos que estuvieron a la vez en el top 5 de EEUU y de UK (indicar también el resultado)



8. Obtener la máxima y mínima posición que ocuparon los discos de Pink Floyd en EEUU y en UK (indicar también el resultado) empleando los comandos de LATIN PIG

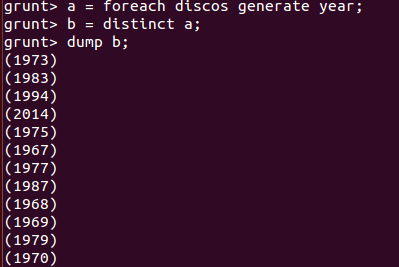


9. Explica con tus propias palabras lo que se desea obtener con los siguientes comandos e indica el resultado obtenido.

a = foreach discos generate anio; //guardo en la variable a los años del disco

b = distinct a; //guardo en b los distintos años que están en a, sin repeticion

dump b; //Muestro el resultado de b



Lo que se consigue es obtener todos los años en los que se publicó un disco, sin repetir el año. Esto último se consigue con la función distinct.

10. (opcional) Empleando UDFs extrae información útil de la discografía.