Práctica 2 - PIG Map Reduce

Vamos a trabajar con el fichero passwd:

- head /etc/passwd

Apache Pig puede ejecutarse en dos modos:

1. **Modo Interactivo (Grunt Shell):** Si ejecutas simplemente pig en la línea de comandos sin ningún archivo de script como argumento, entrarás en el shell interactivo de Pig (conocido como Grunt) donde puedes escribir y ejecutar comandos de Pig línea por línea.
    a. Ejecutar **pig -X**: Este comando se utiliza para iniciar Apache Pig en un modo especial que es útil para la depuración y el diagnóstico. El -X es un indicador para ejecutar Pig en un modo de depuración extendido. Dependiendo del argumento que le siga a -X, puedes activar diferentes tipos de depuración o información de diagnóstico. Por ejemplo, pig -Xlocal ejecuta Pig en un modo especial de depuración para el modo local.
2. **Modo de Script**: Si ejecutas pig seguido de un nombre de archivo (por ejemplo, pig myscript.pig), Pig ejecutará el script contenido en ese archivo.

- pig

Vamos al directorio raíz:

- cd hdfs:///
- ls

Ahora cambiamos al directorio cloudera y creamos un nuevo directorio donde copiaremos el archivo passwd:

- cd /user/cloudera
- ls
- mkdir pwd
- cd pwd
- copyFromLocal /etc/passwd passwd

Cargamos el contenido de passwd en una variable, indicando el separador de campos:

- passwd = LOAD 'passwd' USING PigStorage(':') AS (user:chararray, passwd:chararray, uid:int, gid:int, userinfo:chararray, home:chararray, shell:chararray);

Mostramos el contenido devuelto (se verá cómo lo organiza por tuplas):

- DUMP passwd;

Agrupamos el resultado anterior por el último campo (shell):

- grp_shell = GROUP passwd BY shell;

Y lo mostramos por pantalla:

- DUMP grp_shell;

Finalmente, contamos el número de apariciones de cada shell:

- count = FOREACH grp_shell GENERATE group, COUNT(passwd);
- DUMP count;

—----------
(/bin/bash,10)
(/bin/sync,1)
(/bin/false,2)
(/sbin/halt,1)
(/sbin/nologin,37)
(/sbin/shutdown,1)
—----------