La finalidad de esta práctica volcar datos desde Mysql a HDFS y viceversa con Apache Sqoop.

Vete haciendo capturas de pantalla de todos los pasos que vayas dando así como su resultado, acompañándolas de comentarios descriptivos de los mismos.

APARTADO A

INTRODUCCIÓN

Realizaremos la práctica utilizando el servidor de MySql que hemos instalado en una máquina EC2 de AWS en la práctica anterior.

CONTENIDO

1.- En la carpeta del usuario `maria_dev` en HDFS crea una subcarpeta llamada **`sqoop`** donde guardaremos los archivos de esta práctica.

2.- Importa con Sqoop las tres tablas que creamos en MySql en la práctica anterior.

3.- Importa, dejándola en un solo archivo, todos los datos de los empleados adjuntándoles a cada uno toda la información de su departamento.

4. Muestra en HDFS la ubicación y contenido de los ficheros resultantes.

APARTADO B

INTRODUCCIÓN

En este práctica trabajaremos con los ficheros del *dataset* de Movielens de la práctica 3.

CONTENIDO

1.- En la carpeta del usuario `maria_dev` en HDFS crea una subcarpeta llamada **`movielens`** donde guardaremos los archivos `u.data`, `u.user` y `u.item`.

2.- Utilizando PIG, al archivo `u.user` quítale la última columna con el código postal. Guarda el resultado en el archivo **`u.user2`**.

3.- Utilizando PIG, del archivo `u.item` quédate solamente con las dos primeras columnas (id y título). Posteriormente de la columna título extrae el año y guárdala en una nueva columna (`anio`). En la columna título ha de quedar exclusivamente el título sin el año. Guarda el resultado en un archivo llamado **`u.item2`**

APARTADO C

INTRODUCCIÓN

Realizaremos la práctica utilizando el servidor de MySql que hemos instalado en una máquina EC2 de AWS en la práctica anterior.

CONTENIDO

1.- Crea en tu servidor MySql de AWS una nueva base de datos llamada **`movielens.`**

2.- En dicha base de datos crea tres tablas (**`usuarios`**, **`votos`** y **`peliculas`**) con la estructura adecuada para almacenar los ficheros **`u.data`**, **`u.user2`** y **`u.item2.`** Utiliza sentencias `CREATE TABLE.` Crea los índices y relaciones entre las tres tablas.

3.- Utilizando SQOOP, exporta los tres ficheros de HDFS a sus tablas.

4.- Comprueba con sentencias SELECT que los ficheros se importaron correctamente.

APARTADO D

INTRODUCCIÓN

- Realiza las consultas de esta práctica utilizando DBeaver. Muestra la consulta y una captura donde se vea la salida de las mismas, al menos parcialmente.
- Ten en cuenta que la fecha de la votación está en formato `TIMESTAMP.`

CONTENIDO

1. Top 10 películas más votadas de todos los tiempos (número de votos, no el valor de este).
2. Películas con nota media ≥ 4.5 y al menos 100 valoraciones.
3. Usuarios que han dado más de 300 valoraciones y su nota media.
4. Año con más películas votadas (por número total de votos).
5. Las 5 películas más "*polarizadas*" (mayor desviación estándar, con valoraciones muy extremas) con al menos 50 votos. (Investiga qué función de SQL de da la desviación estándar).
6. Usuarios cuya nota media es menor que la nota media global.
7. Películas que han recibido al menos una valoración de 1 y una de 5 (las más divididas).
8. Top 10 usuarios más activos en 1997 (por número de valoraciones ese año).
9. Películas estrenadas después de 1995 con mejor nota media que "Toy Story (1995).
10. Usuarios que han valorado todas las películas estrenadas en 1993.
11. Evolución mensual del número de valoraciones en 1998.

12. Las 5 películas con mayor aumento de popularidad (comparar 1997 vs 1998).
13. Usuarios que han valorado más películas que la media de su género.
14. Películas que nadie ha valorado con 3 estrellas (solo 1,2,4,5).
15. Ranking de días de la semana con más actividad (lunes, martes…).
16. Usuarios que han dado su primera y última valoración con diferencia > 6 meses.
17. Las 10 películas con mayor ratio 5-estrellas / total valoraciones.
18. UPDATE: Aumenta en 1 año la edad de todos los usuarios (simulación de paso del tiempo).
19. INSERT: Añade una nueva película ficticia estrenada hoy.
20. DELETE: Elimina todas las valoraciones anteriores a 1997.