Tarea Programada #2

- La tarea se debe entregar en el Github según el profesor haya asignado los grupos.
- Los documentos deben ser entregados en el TEC Digital según el profesor haya asignado los archivos.
- Toda tarea debe ser defendida ante el profesor, de tal manera todos los estudiantes deben poder explicar la solución satisfactoriamente.
- ¡Buena Suerte!

Análisis de Musica

La tarea consiste en 3 partes básicas.

1. Un Loader que inserte datos en un almacenamiento que pueda ser leído por Hadoop.
2. Un analizador de Spark que tome los datos en el HDFS y haga el análisis.
3. Una aplicación Web que lea de la base de datos de MariaDB y retorne los resultados.

Loader

Deben bajar el siguiente dataset que es de libre uso:

https://www.kaggle.com/datasets/gabrielkahen/music-listening-data-500k-users

Deben cargar los data sets en el hdfs de Haddop, el cargador debe de tomar los datos e insertarlos en el formato que ustedes decidan, no debe ser el mismo en el que viene el Dataset, debe ser lo que necesiten para poder hacer el análisis.

Análisis

Una vez que tengan los datos de las páginas web, debemos hacer el análisis, usando Spark leyendo del HDFS de Hadoop, se debe poder analizar:

Popularidad y frecuencias básicas

1. Top 20 artistas en general — contar cuántos usuarios incluyen cada artista; mostrar el top 20 y su porcentaje de participación en el total de las reproducciones.
2. Top 20 canciones en general — igual que los artistas, solo que para canciones.
3. Top 20 álbumes en general — igual que los artistas, solo que para albunes.
4. Cuántos usuarios comparten el mismo artista #1 — contar usuarios por su artista principal y reportar la moda y su frecuencia.
5. Distribución de menciones por artista — histograma de veces que aparece cada artista; reportar media, mediana, desviación estándar.
6. Participación del long tail — calcular qué porcentaje de artistas acumula el 80% de las menciones .

Conteos simples por usuario

7. Ítems por usuario — contar cuántos artistas/canciones/álbumes lista cada usuario; reportar media y mediana.
8. Artistas únicos en el conjunto — número total de artistas, canciones y álbumes distintos.

9. Usuarios con listas top-3 idénticas — contar duplicados de top-10 y mostrar las duplicaciones más comunes.

10. Usuarios con gustos muy concentrados — contar usuarios cuyo top 5 pertenece todo al mismo artista.

Coocurrencia y posiciones

11. Pares de artistas más frecuentes — contar cuántas veces dos artistas aparecen juntos en la misma lista de usuario; mostrar top 50 pares.

12. Combinaciones de 3 artistas frecuentes — contar tripletas frecuentes.

13. Solapamiento artista-canción — contar cuántas veces la canción más escuchada de un usuario pertenece al artista más escuchado.

14. Posición promedio por artista — para cada artista, calcular la posición media en las listas de los usuarios que lo incluyen.

15. Frecuencia de que el #1 esté en el top 5 global — proporción de usuarios cuyo artista #1 también figura entre los 5 más populares globales.

16. Estabilidad de posiciones — contar usuarios que tienen el mismo artista en las posiciones #1 y #2.

Comparaciones simples

18. Top artistas entre oyentes— definir oyentes que tienen más de 40 canciones y contar sus artistas principales.

19. Popularidad cruzada entre listas — contar artistas que aparecen con más frecuencia en la lista de canciones vs. la de artistas; reportar la diferencia de conteos.

20. Artistas más diversos — contar cuántos usuarios distintos listan cada artista y cuántas canciones distintas tiene cada artista en el conjunto.

Calidad

21. Conteo de datos faltantes — contar usuarios con campos faltantes o con menos ítems de los esperados.

22. Usuarios atípicos — contar usuarios con recuentos de ítems extremadamente altos o bajos, que están en el percentil 99.

23. Artistas de baja cobertura — contar cuántos artistas aparecen menos de 5 veces.

Los resultados de todos estos análisis deben insertarse en MariaDB/MySQL, el diseño de la base de datos lo deben de hacer ustedes.

Noten que los datos deben estar ya calculados, no se pueden calcular en MariaDB, esta solo debe tener los resultados.

Visualizador

Se debe crear una página web, en donde se puedan Visualizar los datos, pueden hacer cuantas páginas sean necesarias con gráficos y tablas para poder presentar la información del punto 2.

Todas las estadísticas se tienen que mostrar, la página puede ser sencilla, no debe ser complicada o tener drill downs, solo se deben poder mostrar los resultados calculados.

Además, cada grupo debe de hacer 3 conclusiones de los datos que obtuvo.