Tarea Programada #2

- La tarea debe entregarse al profesor por medio del TEC Digital antes del día y la hora convenida.
- La tarea debe contener lo siguiente:
    a. Fuentes, todo el código necesario para ejecutar la tarea. El código debe estar debidamente documentado.
    b. Documentación, incluyendo al menos:
        i. Documentación en el código
        ii. Explicación del diseño y Arquitectura
        iii. Manual de Usuario
    c. Todo debe estar contenido en un Zip file que sea <INICIAL><APELLIDO>, ejemplo EHERNANDEZ.zip, si hay más de un miembro en el equipo, separen los nombres con UNDERSCORE.
- Toda tarea debe ser defendida ante el profesor, de tal manera todos los estudiantes deben poder explicar la solución satisfactoriamente.
- ¡Buena Suerte!

A Evaluar	**Puntos**	**Nota**
Documentación	5	
Aplicación JAVA que conecta a Twitter	20	
Jobs de Map – Reduce (Hadoop)	40	
Aplicación Web	15	
Gráficos	10	
Estabilidad y Completitud	10	
Total	**100**	
Tipos especiales	10	
Total	**110**	

Análisis de Twitter

La tarea consiste de tres partes, la primera es hacer una aplicación en JAVA que se conecte al API de twitter y baje información de los tweets, la segunda debe analizar la información de los tweets y debe insertar el resultado a una base de datos MySQL y la tercera parte es realizar una aplicación Web en JAVA/Javascript que muestre y visualice los datos.

Conectar a Twitter

La aplicación JAVA debe de conectarse al API de Twitter, específicamente al public streamming, y debe de bajar todos los tweets acerca de los siguientes temas.

1. #2030NOW
2. #women
3. #costarica
4. #puravida
5. #MakeAmericaGreatAgain
6. #Trumprussia
7. #RecycleReuse
8. #TraficoCR

Todos los tweets deben de guardarse en un archivo en el hdfs de hadoop, en el formato que prefieran.

La aplicación debe de correr por unos cuantos días para poder obtener suficientes datos para hacer las pruebas, los archivos también deben guardarse para la presentación.

En la siguiente página pueden encontrar la información del API de Twitter.

https://dev.twitter.com/streaming/public

Hadoop

Hadoop debe de tomar la información de los tweets y debe poder sacar los siguientes resultados:

1. Número de usuarios que se recibieron para cada tema
2. Número de tweets recibidos para cada tema
3. Las 10 palabras que más se repitieron en los tweets para cada tema, con su debido total (deben limpiar la información para no tomar en cuenta preposiciones)
4. Los otros 10 hashtags más populares que contiene el tweet para cada uno de los hashtags de la primera parte, con sus respectivos totales.
5. La distribución de tiempo de los tweets hechos por hora, para cada uno de los tags.
6. Para cada hashtag, cuantos tweets incluían alguno de los otros hashtags de la lista de la primera parte.

Este análisis se debe hacer con Map-Reduce usando Hadoop, los resultados deben de insertarse a una Base de Datos MySQL, la cual deben diseñar.

Deben poder hacer un job en el servidor que corra los Map-Reduce automáticamente.

Visualizador

Se debe hacer una aplicación Web, donde se puedan visualizar los resultados con una tabla, usando paginación. Y además, se debe visualizar la información en un gráfico, usando la biblioteca d3js, de tal manera que conforme se filtre la información cambie el gráfico.

Los gráficos y el diseño de la página web quedan a discreción.

La información de la biblioteca se puede encontrar en https://d3js.org

Puntos Extra

1. Hacer que los Jobs de Map – Reduce usen tipos complejos hechos por ustedes y no los tipos que ya Hadoop provee. (10 puntos)