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 Dropbox 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!

Wikipedia Análisis

La tarea consiste en 3 partes básicas.

1. Un Web Crawler 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.

Web Crawler

El Web Crawler debe conectarse a Wikipedia y navegar recursivamente hasta un máximo n (el límite lo definen ustedes) y debe de insertar la información necesaria para los análisis.

El formato con el cual se guarden los datos y el filesystem queda a su criterio. Tomen en cuenta que dependiendo del formato en el que el Web Crawler inserte los datos puede ser más o menos difícil hacer el análisis.

Además, el proceso de procesar e insertar estos datos puede tomar tiempo así que no lo dejen para el último minuto. Esto debe ser la primera parte de la tarea. Mínimo para el file es 1GB.

Lo que se debe guardar será:

1. El Texto ya sea completo o en palabras.
    a. En el caso de las palabras, tienen que remover todo lo que no es una palabra, como por ejemplo (el, de, los, las, un, una, preposiciones, números, etc.)
    b. Vamos a hacer análisis con palabras individuales, dobles palabras y triples palabras, pueden usar solo una lista o tener una para cada análisis, cómo quieran.
2. La dirección web de la página.
3. El título de la página.
4. La cantidad de ediciones que ha tenido la página por día.
5. Los enlaces que hay en la página hacia otras páginas.

Pueden omitir cualquier dirección en el head de la página y cualquier MIME type que no se HTML, como imágenes, documentos, etc., solo nos concentraremos en texto.

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:

1. Para cada palabra distinta, ¿Cuáles son las páginas que más copias de esta palabra tienen?
2. Para cada palabra set de 2 palabras, ¿Cuáles son las páginas que más copias de este set de palabras tiene?
3. Para cada palabra set de 3 palabras, ¿Cuáles son las páginas que más copias de este set de palabras tiene?
4. ¿Cuáles páginas tienen más sets de 2 palabras coincidentes con una página dada?

5. ¿Cuáles páginas tienen más sets de 3 palabras coincidentes con una página dada?
6. Para cada página, ¿Cuál es el set de palabras distintas y cuantas hay de cada una?
7. Para cada página ¿cuántos links distintos aparecen en cada página?
8. ¿Cuál es el porcentaje que cada palabra distinta en el texto total de la página?
9. Hacer un grafo con cómo los enlaces se conectan con otras páginas, de esta manera pueden tener los resultados de cuáles son los tópicos que más interconectados están con otras páginas.
10. Agregar 2 análisis distintos que ustedes crean pertinente.

Los resultados de todos estos análisis deben insertarse en MariaDB, 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.

Buscador

Se debe crear una página web, en donde se puedan hacer consultas a la base de datos de María DB, estas se van a devolver dependiendo de las palabras que se busquen, y se ordenan de acuerdo con algún algoritmo que especifiquen usando los resultados.

- En el resultado debe ir:
    a. El título de la página
    b. El enlace de la página original
    c. Todas las estadísticas que se calcularon para esa página en la parte 2.