Sistema de reportes y análisis sobre tendencias en la Web de la ESPOL usando Hadoop para el procesamiento masivo de los datos.

Gallardo Luis, Bermeo Fabricio, Cedeño Vanessa Msc.
Facultad de Ingeniería en Electricidad y Computación
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
{luigalla, fbermeo, vcedeno}@espol.edu.ec

Resumen

Los sistemas de reportes y análisis sobre tendencias son ampliamente utilizados hoy en día gracias a su capacidad de analizar las palabras más usadas por los usuarios en la web, por ejemplo se han usado estos sistemas en las redes sociales, ya que las mismas en los últimos años han tenido gran acogida. En la web de la ESPOL existe una gran cantidad de datos, y no existe una herramienta que permita recolectarlos y realizar un análisis de su contenido, para resolver este problema usamos Hadoop que es una plataforma que nos permite desarrollar aplicaciones que tengan que tratar con grandes cantidades de datos, hasta petabytes. Los programas MapReduce de Hadoop están diseñados para computar grandes volúmenes de datos en paralelo. El ejemplo más claro de esto, lo que va a ser de gran ayuda y que se usa para este tipo de problemas es el WordCount, que lee archivos de texto y cuenta con qué frecuencia ocurren las palabras, como resultado final vamos a tener gráficos estadísticos para analizar las tendencias.

Palabras Claves: *Sistemas de reportes y análisis, Hadoop, WordCount, gráficos estadísticos.*

Abstract

Reporting and analyzing systems about trends are widely used nowadays due to their capacity to analyze the words most used by web users, for example these systems have been used in social networks, since they have been very popular in the last years. There is a lot of data on the ESPOL website, and there is not a tool to collect them and make an analysis of its content; to solve this problem we use Hadoop which is a platform that allows us to develop applications that have to deal with large amounts of data, even with petabytes. Hadoop MapReduce programs are designed to compute large amounts of data in parallel. An example about this and that will be used for such problems is the WordCount. It reads text files and records how often the words occur, as a final result we are going to have statistical graphics to analyze the trends.

Keywords*: reporting and analysis systems, Hadoop, WordCount, statistical graphics.*

1. Introducción

Durante los últimos años, los sistemas de reportes y análisis sobre tendencias han sido herramientas que han contribuido a mejorar la experiencia de los usuarios en las aplicaciones que éstos utilizan, dado que los mismos nos permiten conocer acerca de los temas que son de mayor popularidad en los diferentes sitios.

Numerosas empresas han apostado por el uso de estas herramientas para lograr que sus sitios web poseen un ambiente personalizado, en donde los usuarios puedan observar las palabras más usadas por ellos mismos. Algunos de los sitios web en donde se puede encontrar el uso de estos sistemas son Google, Facebook, Twitter, etc., los cuales permiten ver un análisis estadístico de las palabras más usadas por los usuarios de estos sitios.

El presente artículo introduce una solución informática que permite adaptar un sistema de análisis y reportes basados en el usuario para realizar gráficos estadísticos cuando los estudiantes requieran conocer los temas más populares dentro de la red de la universidad.

El contenido de este artículo se distribuye de la siguiente manera: en la sección 2 se discute la motivación que llevó a escribir el artículo. En la sección 3 se describe el trabajo relacionado, seguida de la sección 4 dedicada a definir conceptos básicos sobre las

tecnologías que son usadas en los sistemas de reportes y análisis. En la sección 5 se describe la metodología empleada para el desarrollo de la solución y las pruebas realizadas. La sección 6 incluye el análisis de los resultados. Finalmente, en la sección 7 damos nuestras conclusiones y en la 8 los agradecimientos.

2. Motivación

Dada la gran cantidad de información que existe en la red social de la ESPOL, y ya que en la actualidad no existe una herramienta que nos permita procesar todos esos datos y con esto llegar a conocer los temas de interés de los estudiantes. El presente trabajo nace como una iniciativa que busca explorar la viabilidad de la aplicación de los sistemas de reportes y análisis sobre tendencias en la web de la universidad, y hallar un escenario para la implementación de este tipo de sistemas dentro de ella. De esta manera construir un trabajo que sirva de base a futuros proyectos dentro de ESPOL.

3. Trabajos relacionados

En el ámbito del internet los sistemas de reportes y análisis sobre tendencias han sido ampliamente usados para el beneficio de sus propios usuarios, que les permite conocer sobre sus propios temas de interés.

Como ejemplo de esto tenemos a Trendsmap [1] que es una interesante herramienta para conocer en tiempo real que es lo que está sucediendo en tu región, cuales son los hashtag más populares y sobre que tendencias se está hablando en Twitter, pero Trendsmap no es una herramienta de tendencias cualquiera, lo caracteriza su tecnología de GeoIP gracias a Google, lo cual nos permite tener cierta precisión sobre la ubicación de un hashtag en una determinada región.

También existe Google Trends [2] que es una herramienta de Google Labs que muestra los términos de búsqueda más populares del pasado reciente. Las gráficas de Google Trends representan con cuánta frecuencia se realiza una búsqueda particular en varias regiones del mundo y en varios idiomas. Una característica adicional de Google Trends es la posibilidad de mostrar noticias relacionadas con el término de búsqueda encima de la gráfica, mostrando cómo afectan los eventos a la popularidad.

Otro ejemplo de estas herramientas es el Lexicon de Facebook [3], que fue usada hace algún tiempo y que por el momento se encuentra deshabilitada, esta es una herramienta para seguir las tendencias lingüísticas en esta red social, consulta el uso de palabras y frases en los muros de perfiles, grupos y eventos. Por ejemplo, puedes introducir "amor, odio" para comparar el uso de estas dos palabras en los muros de Facebook. Por último, cabe agregar que esta aplicación nos permite determinar (o al menos tener una aproximación) del éxito que una campaña viral ha tenido en Facebook.

4. Conceptos y Tecnologías Empleadas

Hadoop [4] es una plataforma que nos permite desarrollar aplicaciones que tengan que tratar con grandes cantidades de datos, hasta petabytes. Se trata de un sub-proyecto de Lucene, un proyecto de Apache que desarrolla software para realizar búsquedas. Esta plataforma es muy útil cuando vamos a realizar proyectos que necesiten de escalabilidad, ya hemos dicho que puede almacenar y procesar petabytes de información.

A su vez, es perfecto para un clúster de servidores, distribuyendo la información entre los nodos, siendo posible disponer de miles de nodos. Al disponer los datos de forma distribuida, la búsqueda se puede realizar muy rápidamente ya que Hadoop puede acceder a ella de forma paralela. Y aunque los datos estén distribuidos, no hay que preocuparse de fallos ya que dispone de un sistema de seguridad.

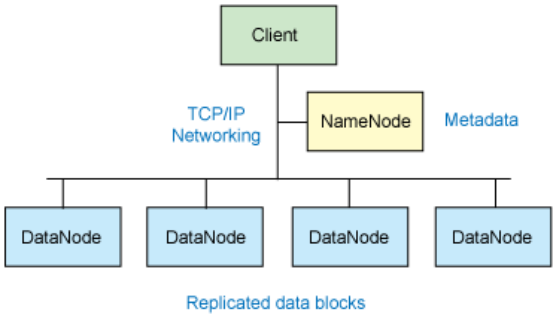

Figura 1. Visión simplificada de un clúster Hadoop

Hadoop consiste básicamente en el Hadoop Common, que proporciona acceso a los sistemas de archivos soportados por éste. El paquete de software 'The Hadoop Common' contiene los archivos .jar y los scripts necesarios para hacer correr Hadoop. El paquete también proporciona código fuente, documentación, y una sección de contribución que incluye proyectos de su Comunidad.

El Hadoop Distributed File System (HDFS) es un sistema de archivos distribuido, escalable y portátil escrito en Java para el framework Hadoop. Cada nodo en una instancia Hadoop típicamente tiene un único nodo de datos; un clúster de datos forma el clúster HDFS (En la Figura 1 nos muestra una visión de un clúster). La situación es típica porque cada nodo no requiere un nodo de datos para estar presente, cada nodo sirve bloques de datos sobre la red usando un protocolo de bloqueo específico para HDFS. El sistema de archivos usa la capa TCP/IP para la comunicación; los clientes usan RPC para comunicarse entre ellos. El HDFS almacena archivos grandes (el tamaño ideal de archivos es de 64 MB), a través de múltiples máquinas, consigue fiabilidad mediante replicada de datos a través de múltiples hosts, y no requiere almacenamiento RAID en ellos. Con el valor de replicación por defecto: 3, los

datos se almacenan en 3 nodos: dos en el mismo rack, y otro en un rack distinto. Los nodos de datos pueden hablar entre ellos para reequilibrar datos, mover copias, y conservar alta la replicación de datos.

MapReduce [5] es un modelo de programación diseñado para procesar grandes volúmenes de datos en paralelo dividiendo el trabajo en un conjunto de tareas independientes. Los programas MapReduce son escritos en un particular estilo influenciado por construcciones de programación funcional, específicamente lenguajes para el procesamiento de listas de datos. En este módulo se explica la naturaleza de este modelo de programación y cómo puede ser utilizado para escribir programas que se ejecutan en el entorno de Hadoop. Estos programas están diseñados para calcular grandes cantidades de datos en forma paralela y para ello es necesario dividir la cantidad de trabajo entre un gran número de máquinas. Este modelo no escalaría a clústeres (cientos o miles de nodos) si a los componentes se les permitiría compartir los datos de forma arbitraria. Los gastos generales de comunicación necesarios para mantener los datos en los nodos sincronizados en todo momento evitarán que el sistema funcione de forma fiable y eficiente a gran escala. Por el contrario, todos los elementos de datos en MapReduce son inmutables, es decir, que no pueden ser actualizados. Si en una tarea de mapeo se cambia una entrada (clave, valor) par, no se refleja en los archivos de entrada, la comunicación se produce únicamente mediante la generación de una nueva salida (clave, valor) pares que después son remitidos por el sistema de Hadoop en la siguiente fase de ejecución. Conceptualmente, los programas de MapReduce transforman listas de elementos de datos de entrada en listas de elementos de datos de salida. Un programa de éste tipo hará esto dos veces, usando dos diferentes lenguajes de procesamiento de listas: Map y Reduce, estos términos son tomados de varios lenguajes de procesamiento de listas, tales como LISP, Scheme o ML. Un ejemplo de esto es el WordCount que es un simple programa MapReduce que fue escrito para determinar cuántas veces aparecen diferentes palabras en un conjunto de archivos.

Una herramienta sencilla Geany [6] es usada para desarrollar el código en java, es un editor de texto ligero basado en Scintilla con características básicas de entorno de desarrollo integrado (IDE). Está disponible para distintos sistemas operativos, como GNU/Linux, Mac OS X, BSD, Solaris y Microsoft Windows. Es distribuido como software libre bajo la Licencia Pública General de GNU.

Además para la realización de cualquier gráfico personalizado tenemos a Adobe Flex [7] que es un marco de trabajo gratuito de código abierto y altamente productivo para crear aplicaciones web, para ordenadores de escritorio y para dispositivos móviles, este herramienta permite crear aplicaciones web y para dispositivos móviles que comparten una base de código común, lo que reduce el tiempo y el coste de creación de aplicaciones y el mantenimiento a largo plazo.

Por último tenemos a la herramienta MySQL [8] que es un sistema de administración de bases de datos (Database Management System, DBMS) para bases de datos relacionales. Fue escrito en C y C++ y destaca por su gran adaptación a diferentes entornos de desarrollo, permitiendo su interactuación con los lenguajes de programación más utilizados como PHP, Perl y Java y su integración en distintos sistemas operativos.

5. Materiales y Método

5.1. Adaptación del Sistema de Reporte y Análisis

La página de facebook de la ESPOL nos ofrece todos los registros de las publicaciones y comentarios de los estudiantes, acerca de los temas de actualidad en la universidad, estos datos nos pueden ayudar a conocer sobre que temas son los más populares, es decir los que más son de objetivo de comentarios. Por ejemplo tenemos un gran universo de palabras del cual vamos a escoger una, a ésta se la analizará con respecto al número de veces que se ha nombrado en la web y se mostrará el respectivo reporte de la misma. Además también se podrá comparar dos o más palabras para ver cual de ellas ha sido de más popularidad entre los usuarios.

El sistema para llegar a mostrar sus resultados finales se lo tuvo que realizar por diferentes etapas y con diferentes herramientas.

Figura 2. Post de la página de facebook de la ESPOL

El procedimiento para obtener el análisis y reporte de las tendencias se centra en la obtención de los todos los comentarios que existan en la página de facebook de la ESPOL, para esto se realizó dos scripts, a continuación se detalla cada uno de ellos. El primer script se trata de la obtención de los URL de todos los post publicados en la página web y el segundo script es

para obtener el contenido de todos los post. Aquí nos conectamos con cada URL del post y obtenemos la fecha, el contenido y los comentarios de cada uno. En la figura 2 tenemos un ejemplo de un post.

Como de la página de facebook de la ESPOL obtenemos una gran cantidad de palabras las cuales se repiten muchas veces, para este tipo de casos usamos el WordCount, para que nos agrupe en una sola palabra y el número de repeticiones de la misma.

El WordCount es una sencilla aplicación que se encarga de analizar un texto, señalando en una lista: las palabras utilizadas y el número de repeticiones existentes, el formato de los datos es: como entrada un archivo de texto plano (.txt) y como salida la palabra que es la clave y el número de repetición que es el valor. La implementación de WordCountMapper, vía el método mapper, toma cada línea del archivo de texto, según lo dispuesto por el TextInputFormat. La línea es dividida en palabras y emite pares clave/valor (palabra, 1). La implementación del WordCountReducer, vía el método reducer, recibe la salida del mapper que contiene todas las palabras existentes en el archivo de entrada y cuenta la frecuencia de ellas emitiendo un nuevo par clave/valor (palabra, número de repetición).

Por último almacenamos todos estos resultados en una base de datos para luego ser leídos y usados para mostrar los reportes estadísticos. La estructura del sistema de reporte y análisis de tendencias, se muestra en la figura 3.

5.2. Pruebas realizadas

Usando el Sistema de reportes y análisis sobre tendencias implementado, se hicieron la comparación de un máximo de 5 palabras usadas por los usuario de la comunidad de la universidad, mostrando cuales fueron las palabras más mencionadas por los estudiantes.

Para obtener un resultado actualizado, podemos seguir paso a paso las pruebas que se fueron realizando hasta obtener los resultados finales. Para eso se realizó lo siguiente:

Se realizaron pruebas del script de obtención de las palabras de la web de la ESPOL, esto abarca la adición de la fecha en que se publicó una palabra, luego de haber realizado esto se probó aplicando el WordCount a las palabras que se obtienen de la web, esto analizará y obtendrá el número de repeticiones de cada palabra.

Seguidamente de haber obtenido todas las palabras con su fecha y con el número de repeticiones se realizó las pruebas correspondientes de la subida de los datos a la base MySQL, además con los datos ya subidos se prueba la conexión con Adobe Flex y ver si se obtienen los datos correctamente. Y finalmente con ésta última herramienta se probará la generación de los gráficos correspondientes.

Todo lo detallado anteriormente fue el plan de pruebas que se realizó desde el comienzo del desarrollo hasta obtener el resultado final del sistema.

El siguiente paso fue analizar que palabras fueron las más mencionadas, para esto debemos ver los resultados del análisis de cada palabra en los gráficos de reportes, además podemos comparar entre palabras que pueden ser sinónimos o palabras completamente diferentes, todos estos escenarios tendrán que estar presentes en la pruebas finales a realizarse para obtener un informe detallado e ir comprobando cual tema es la tendencia actual.

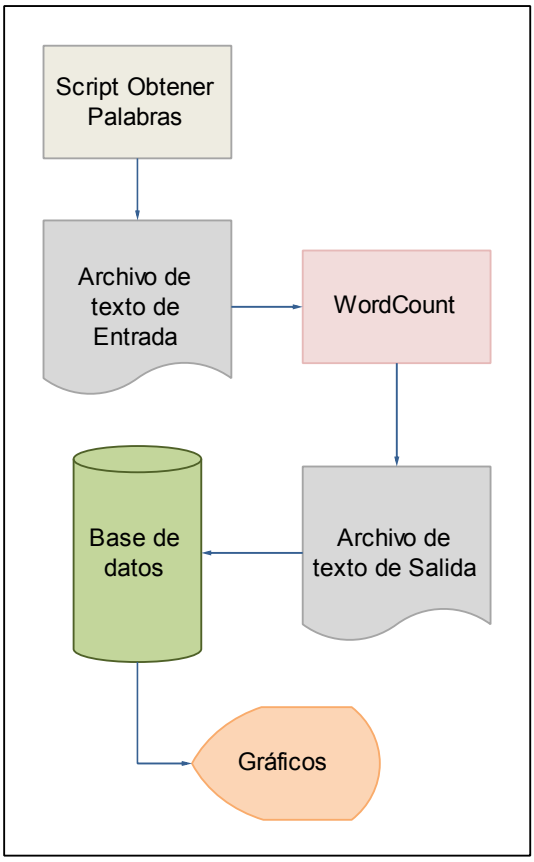

Figura 3. Estructura del sistema

6. Análisis de Resultados

Analizamos los datos de la generación del archivo de texto con todo ese proceso para garantizar un resultado óptimo. Dado esto podemos tomar muestras de ciertas palabras para comprobar la existencia de las mismas y su número de repeticiones, además con el reporte de gráficos podemos tener más claridad, el cual es mostrado a continuación.

En la Figura 4 mostramos la pantalla gráfica en donde ingresaremos las palabras para su correspondiente análisis.

Después de haber ingresado las palabras para su correspondiente búsqueda nos muestra un gráfico de las tendencias de las palabras por año y por mes, esto podemos ver en la Figura 5.

En la Figura 6 nos muestra un gráfico de las tendencias de las palabras de un mes escogido, por ejemplo si escogemos el mes de noviembre del año 2011 nos va a mostrar las tendencias de las palabras en dicho mes.

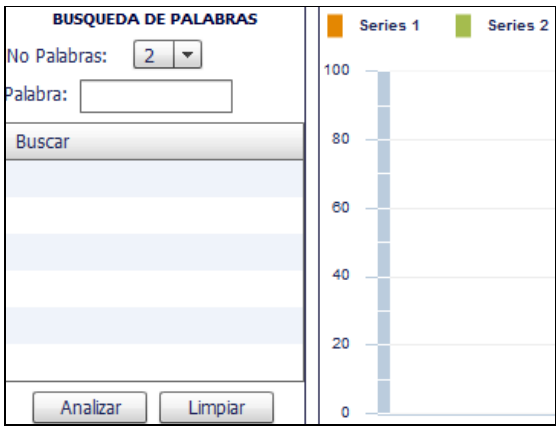

Figura 4. Pantalla de ingreso de las palabras

Figura 5. Gráfico de las tendencias por año y mes

Figura 6. Gráfico de tendencias mensuales

7. Conclusiones

Java nativo de Hadoop permite maximizar el excelente uso de recursos, para obtener el resultado más óptimo, pero sacrifica: facilidad en escritura de código y tiempo empleado en la implementación de la solución

El procesamiento de archivos pequeños (en el orden de los KB) demora la tarea de ejecución debido a que Hadoop está desarrollado para ser más eficiente manipulando archivos de gran tamaño.

El sistema de reportes y análisis nos permitió conocer sobre los temas más comunes en la comunidad de la ESPOL de los últimos meses, ayudándonos a ver en detalle, con gráficos correspondientes que nos muestra las tendencias y estadísticas de éstos temas.

8. Agradecimientos

Este proyecto es el resultado del esfuerzo conjunto de los que formamos el grupo de trabajo. Agradecemos a todas las personas que de una u otra manera colaboraron en la realización de este trabajo, en especial a la Ing. Vanessa Cedeño Directora de Proyecto y al Ing. Lenín Freire Miembro Principal.

9. Referencias

[1] Alejandro, “Trendsmap la herramienta de tendencias en tiempo real para twitter con GeoIP”, http://www.debubuntu.com/trendsmap-la-herramienta-de-tendencias-en-tiempo-real-para-twitter-con-geoip/, Septiembre 2011

[2] Wikipedia, Google Trends, http://es.wikipedia.org/wiki/Google_Trends , Agosto 2012

[3] Milagros, Facebook Lexicon compara palabras en Facebook, http://www.chicaseo.com/facebook-lexicon/ , Junio 2007

[4] Wikipedia, Hadoop, http://es.wikipedia.org/wiki/Hadoop, Agosto 2011

[5] Yahoo, MapReduce, http://developer.yahoo.com/hadoop/tutorial/module4.html , Octubre 2011

[6] Wikipedia, Geany, http://es.wikipedia.org/wiki/Geany , Octubre 2005

[7] Adobe, What is Flex?, http://www.adobe.com/products/flex.html, Noviembre 2011

[8] José Manuel Pérez, Que es MySQL?, http://www.esepestudio.com/articulo/desarrollo-web/bases-de-datos-mysql/Que-es-MySQL.htm, Agosto 2005