ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"MINERÍA DE LOGS DE UNA APLICACIÓN SOCIAL MULTIUSUARIO EN LÍNEA"

TESIS DE GRADO

Previa a la obtención del Título de:

INGENIERO EN CIENCIAS COMPUTACIONALES
Especialización SISTEMAS DE INFORMACIÓN

INGENIERO EN CIENCIAS COMPUTACIONALES
Especialización SISTEMAS MULTIMEDIA

Presentado por:

Carlos Andrés Granda Bermúdez

Luis Elicio Loaiza Pacheco

Guayaquil, Ecuador

2009

AGRADECIMIENTO

ING. CRISTINA ABAD ROBALINO Directora de Tesis, por su ayuda y colaboración para la realización de este trabajo.

ING. SERGIO FLORES Decano de la FIEC, por su constante apoyo y dirección.

AMIGOS Y FAMILIARES.

DEDICATORIA

A MIS PADRES

MI ABUELA Y TIAS

MIS HERMANOS

MI ESPOSA E HIJOS

Carlos Andrés Granda

TRIBUNAL GRADUACIÓN

Msc. Cristina Abad Robalino
DIRECTOR DE TESIS

Ing. Sergio Flores
DECANO FIEC

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponde exclusivamente; y el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral".

Carlos Andrés Granda Bermúdez

Luis Elicio Loaiza Pacheco

RESUMEN

El núcleo del proyecto es realizar un procesamiento masivo de bitácoras de gran tamaño, que son producto de las actividades generadas por usuarios de una aplicación social en un ambiente virtual. La aplicación adapta publicidad inmersa en el mundo virtual con el fin de tener un mayor protagonismo. El procesamiento extrae información de estos registros que sirven para proveer de conocimiento de potenciales clientes, a las empresas o marcas presentes en la aplicación.

Con el fin de que el proceso de minería sea eficiente y no consuma recursos propios, se usa la plataforma de procesamiento distribuido Hadoop, y los servicios que provee Amazon, de computación en la nube (Cloud Computing), para el desarrollo de este caso de estudio.

Para la realización del proyecto, se ha utilizado el API de Facebook y así obtener datos de usuarios reales ya que la aplicación no está en ejecución real. Se adapta a su vez un formato de logs o bitácoras preestablecido. También se hace uso de datos de empresas que especifican características de usuarios de su interés.

Se hace uso de Java y de PIG como herramienta que proporciona un lenguaje de alto nivel (PIG Latin), para fácilmente procesar las fuentes de datos. Esta herramienta administra los procesos Map-Reduce necesarios para completar la tarea de minería.

Al final de este procesamiento masivo de datos se logra obtener información con respecto a las características de usuarios o mercado objetivo antes especificado por cada empresa o marca en la aplicación. Esta información indica cantidad de usuarios

que interactuaron con la marca, como lo hicieron, como la mayoría de usuarios se comporta con cada marca, y tendencias de interacción con objetos relacionados a marcas en el mundo virtual.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

INTRODUCCION

El presente proyecto realiza minería de logs (bitácoras) de gran tamaño, producto de las actividades generadas por usuarios simultáneamente conectados a una aplicación social que permite interactuar con objetos relacionados a marcas o productos reales en un ambiente virtual. La minería de logs es un proceso que permite extraer información que reside de manera implícita en los datos almacenados en este tipo de registros.

Empresas están siempre en busca de información con respecto a sus clientes y la forma de poder obtener nuevos clientes. Los nuevos clientes usualmente forman parte de los mercados objetivos pero es probable que aún no conozcan el producto o marca ofertada o sean clientes de la competencia. Esta es la razón por la que se utilizan todos los medios posibles para continuamente monitorear la reacción de los usuarios con respecto a sus productos o marcas y plantear efectivas estrategias de marketing, e-marketing y publicidad.

Hoy en día el "In-game Advertising" [1] o "publicidad integrada en juegos" se convierte en una estrategia de marketing que trae grandes beneficios, entre los cuales tenemos: mayor recordación, mayor exposición por parte del consumidor, mejor comunicación de los atributos de sus productos o servicios. Esto permite que los juegos generen junto con la actividad de los usuarios, contenido valioso para ser analizado dentro de un estudio de mercado real.

CAPÍTULO 1

Generalidades

1.1 Análisis del problema

Los mercados objetivos de una empresa siempre están definidos con características claramente identificables. Varias de estas características comprenden: sexo, edad, gustos, grupos o comunidades a los que pertenecen y como se relacionan con otros. A su vez para las empresas es muy importante que los medios que se usen para este objetivo sean de alcance masivo. Las aplicaciones o juegos dentro de redes sociales se han convertido en un medio propicio para obtener información que describa el comportamiento del mercado en relación a un producto o marca.

Las aplicaciones sociales contienen bitácoras de actividades, que registran gran cantidad de datos producto de las actividades que sus usuarios generan espontáneamente. Realizar un análisis de estas bitácoras a través de minería trae como resultado información que permite conocer rasgos de comportamiento y preferencia de los usuarios. Si estas bitácoras son vinculadas directamente hacia una marca o producto específico, es posible obtener la información que tanto buscan las empresas. Los logs (bitácoras) de actividades almacenan a diario gran cantidad de información, alcanzando tamaños mensuales promedio de hasta 100 GB.

Para realizar el procesamiento de estas fuentes de datos es vital hacerlo de la manera más eficiente posible.

En la actualidad, existe un framework de procesamiento masivo y escalable de datos llamado Hadoop[2] el cual tiene mucha acogida sobre todo en empresas de la Web 2.0. Este framework proporciona una estructura escalable para brindar el beneficio de procesar más eficientemente gran cantidad de datos como los que existirían en un log de actividades.

1.2 Alcance

El objetivo del proyecto es realizar minería de logs de una aplicación social multiusuario en linea, a través del framework Hadoop que soporta procesamiento masivo y distribuido de datos. Para entregar como resultado indicadores identificables referentes a los usuarios de la aplicación, las marcas y sus mercados objetivos.

Con estos indicadores se obtienen las preferencias y comportamientos de los usuarios dentro de la aplicación con respecto al producto o marca ofertada por una empresa. De esta manera se busca retroalimentar a la empresa con resultados que muestran el éxito o rechazo de su producto o marca dentro de su mercado objetivo.

Finalmente el objetivo más ambicioso del proyecto, es poder inferir, luego del análisis de resultados obtenidos del proceso anterior, la existencia de mercados potencialmente interesados en los productos o marcas y que la empresa no había identificado con anterioridad.

CAPÍTULO 2

Diseño de la solución

2.1 Datos de entrada

Para este caso de estudio específico se consideró una aplicación social multiusuario en línea enlazada con los datos de usuario disponibles en Facebook. Esta aplicación utiliza el concepto de publicidad inmersa en el juego. La aplicación está actualmente en etapa de desarrollo por parte de los miembros que realizan este proyecto.

La aplicación anteriormente mencionada posee un middleware (Socket Server), el cual al detectar alguna actividad de usuario en línea almacena los datos referentes a la misma agregando una entrada al log de actividades. Este log de actividades es el que tomaremos como la fuente principal de datos de entrada para llevar a cabo la minería y obtener la información planteada.

2.2 Framework para procesamiento distribuido

El framework o entorno de desarrollo utilizado es Apache Hadoop en conjunto con otras herramientas de apoyo como Pig[3]. Esta plataforma facilita el trabajo de programación y, para este caso especifico, ayuda a obtener los grupos de usuarios que recogen ciertas características.

Pig es una herramienta para el análisis de grandes fuentes de datos. Consiste en un lenguaje de alto nivel para expresar programas de flujo de datos, conjuntamente con una infraestructura para la evaluación de estos programas. La característica sobresaliente de los programas de Pig es que su estructura permite paralelizar procesos, que a su vez manejan fuentes de datos de gran tamaño.

La capa de infraestructura de Pig consiste en un compilador que produce secuencias de programas Map/Reduce [4], para los cuales ya existe implementaciones en paralelo (Hadoop). El lenguaje de Pig es un lenguaje textual llamado Pig Latin, el cual tiene las siguientes características [5]:

- **Fácil programación**. Las tareas complejas que comprenden múltiples transformaciones de datos relacionados entre sí, son explícitamente codificadas como secuencias de flujo de datos, haciéndolas fáciles de escribir, entender y mantener.
- **Oportunidades de optimización**. La forma en que las tareas son codificadas permite que el sistema optimice automáticamente su ejecución, permitiendo al usuario enfocarse en la semántica en lugar de la eficiencia.

- **Extensibilidad**. Usuarios pueden crear funciones especializadas para procesamiento específico.

2.3 Plataforma de Cloud Computing

Otro concepto clave para realizar este proyecto "Cloud Computing" y la herramienta para este propósito son los Amazon Web Services[6], específicamente el Elastic Computing Cloud (EC2) [7] para levantar un cluster Hadoop y el Simple Storage Service (S3) [8] para el almacenamiento de los datos de entrada y de salida. Esta infraestructura distribuida ayuda a hacer más eficiente el sistema permitiendo escalabilidad y eficiencia para procesar la gran cantidad de información que se prevé manejar.

2.4 Flujo de procesos

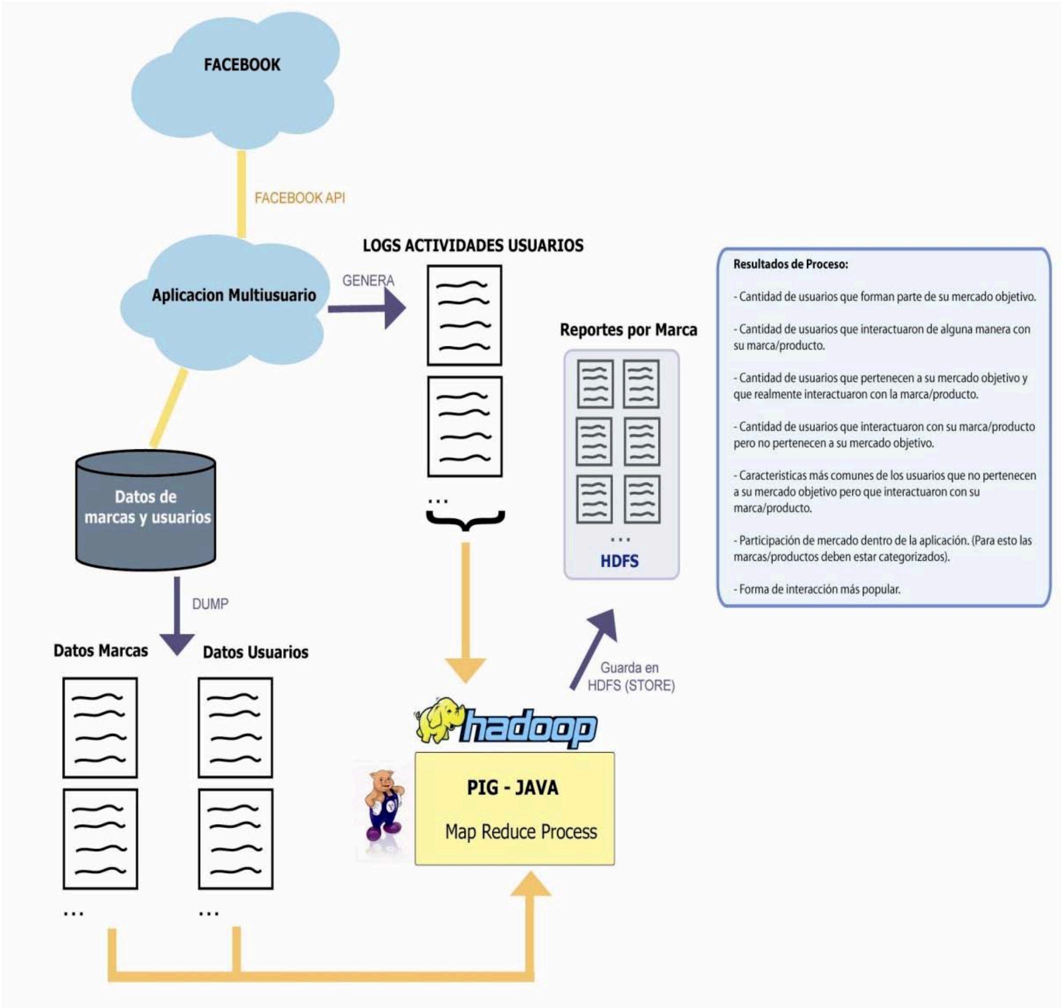


Figura 1. Flujo de procesos

La Fig. 1 muestra el diseño de flujo de procesos. Se obtienen primero los datasets de entrada (logs de actividades y dump de usuarios y marcas), tomando en cuenta los formatos que se establecieron en etapas previas para la concepción de la idea. El archivo de marcas incluye condiciones de mercado que las empresas definen con anterioridad al momento de aceptar ser parte de la aplicación.

El archivo de datos de usuarios fue generado a partir de llamadas al API de Facebook[9], que nos permite obtener valores reales de una red social.

Estos archivos son los que servirán para alimentar el proceso masivo de minería vinculando todos los datos respectivos de cada usuario.

Se procesan los archivos con los datos de las marcas/productos. Usando Pig embebido en JAVA sobre la infraestructura de Hadoop se obtienen las características que las empresas requieren en los usuarios de la aplicación y luego utilizar esta información junto con el archivo de datos de usuarios para generar los grupos objetivos.

Los resultados previos obtenidos más los data sets fuente, son utilizados para generar mercados identificables dentro de la aplicación, y de acuerdo al análisis de actividades en los logs, verificar la existencia de otros mercados potenciales que podrían ser de interés para las marcas. La salida de este proceso genera un reporte que contiene los indicadores por cada marca o empresa.

CAPÍTULO 3

Implementación

En este capítulo se describe la estructura de cada una de las fuentes de entrada para este caso de estudio específico. Los formatos fueron establecidos de acuerdo a las necesidades del proyecto y la estructura que maneja la aplicación de apoyo del proyecto. Estos formatos son validos para implementaciones con otras aplicaciones como también puede variar dependiendo del contexto en que se desee desarrollar.

Como entradas para obtener los resultados planteados contamos con los siguientes data sets:

- Información de los usuarios
- Información de las marcas/empresas
- Constantes definidas
- Log de actividades de la aplicación

3.1 Información de Usuarios

Es un documento de texto (.txt) que contiene datos básicos de los usuarios que utilizan la aplicación. Los campos están separados por tabulaciones con el siguiente formato:

Tabla 1. Formato para datos de Usuarios

Id Usuario	Nombre	Fecha Nacimiento	Edad	Sexo
125	George Enrique Reyes Tomalá	July 4, 1985	23	male

Id Usuario: Código numérico de identificación única de usuario.

Nombre: Nombres y Apellidos de usuario (no hay un formato específico).

Fecha Nacimiento:

Fecha de nacimiento del usuario, en el formato <Mes día, año>, en ciertos casos los usuarios no proporcionaron el año de nacimiento, por lo cual estos usuarios no formaron parte de los análisis realizados).

Sexo: Masculino (male) o Femenino (female).

3.2 Información de marcas/empresas

Es un documento de texto que proporciona datos básicos de las empresas presentes en la aplicación y características de sus marcas. Los campos están separados por tabulaciones y el documento presenta el siguiente formato:

Tabla 2. Formato de datos de empresas

Id Marca Producto	Nombre	Empresa	Características Mercado Objetivo
12	HotWheels	Mattel	edad<=18 and edad>=25;sexo=masculino; claves=autos,juguetes

Id Marca Producto: Código numérico de identificación único de cada marca o producto por empresa.

Nombre: Nombre de la Marca o Producto que se ha ofertado.

Empresa: Nombre de la Empresa responsable del Producto o Marca.

Características Mercado Objetivo:

String que recoge cada una de las características que forman su mercado objetivo. Posee el siguiente formato:

caracteristica1=valor1,valor2;caracateristica2=valor3,valor4;

3.3 Constantes definidas

Dentro de las entradas del log de actividades se encuentran dos clases de códigos que establecimos para identificar los tipos de objetos y las actividades.

Actividades: palabras clave para identificar actividades específicas de usuario dentro de la aplicación.

- iniciar (el usuario ha iniciado sesión).
- comentar (el usuario ha dejado su opinión en comentario con respecto a algún objeto o evento).
- usar (el usuario está utilizando alguna marca en sus pertenencias).
- personalizar (utilizar la marca como distintivo en objetos que le pertenecen, camisetas, autos, etc.).
- comprar (un objeto auspiciado por una marca fue adquirido por un usuario).

- vender (un usuario ha vendido un producto de una marca especifica).
- finalizar (El usuario ha cerrado sesión o no se ha registrado actividad por un largo tiempo).

Tipos: Código numérico para identificar los tipos de objetos involucrados en cada tipo de actividad. Puede ser un producto involucrado en una transacción de compra/venta; un lugar respecto al cual se dejo un comentario o se dio un voto (ranking).

3.4 Log de actividades de la aplicación

Inicialmente es necesario identificar un formato de log de actividades del sistema, que contenga los datos necesarios para poder obtener la información que nos hemos planteado encontrar.

Dado que la aplicación aún no está en producción, generamos nuestros logs artificialmente siguiendo el formato anteriormente establecido por el grupo de trabajo involucrado. La Tabla 3 muestra el formato del log generado. Se prevé que este será el formato que generará la aplicación cuando esté en producción, pero si llegara a determinar que es conveniente cambiarlo por otro formato, esto no requeriría cambios sustanciales en el módulo de minería de logs.

Tabla 3. Formato log de actividades

DateTime	IdUsuario	Actividad	Tipo	Marca
12	126	Personalizar	1	12

DateTime: Fecha y hora en que ocurrió el evento.

IdUsuario: Código de identificación único para cada usuario dentro de la aplicación.

Actividad: De acuerdo a las formas de interacción disponibles en la aplicación, identificamos actividades y las agrupamos en las siguientes palabras claves: iniciar, usar, personalizar, comprar, vender, finalizar.

Tipo: Es un número entero que identifica el tipo de objeto al que se hace referencia. Los códigos válidos son números del 1 al 4.

Marca: Código (numérico/alfanumérico) que identifica el nombre de la marca involucrado.

Los tres tipos de archivos especificados anteriormente son alimentados al sistema distribuido usando "Cloud Computing". Se transfieren estos archivos hacia el Hadoop Distributed File System (HDFS) para luego ser procesados por un programa en Java que también se transfiere al nodo maestro del servicio.

Todo el proceso de minería es ejecutado en un clúster de los servicios de Amazon EC2, con una imagen Amazon Machine Image (AMI) Fedora de Cloudera[10] que soporta la versión de PIG 0.2 y JAVA 1.6.0.10.

A continuación se presenta la porción más importante del código (en lenguaje PigLatin) de la aplicación. Este segmenta los mercados de cada marca o empresa y filtra los datos. Genera los indicadores para los reportes finales por cada marca. Por facilidad de ejecución y manejo de variables, este código es embebido en Java. El compilador de Pig se encarga de generar los procesos Map y Reduce necesarios para paralelizar las tareas de tal manera que se puedan ejecutar de manera distribuida.

```
/*
 * Ejemplo de filtro para obtener los usuarios que pertenecen al mercado objetivo de
 * una marca específica.
 */
log = load 'proyecto/logSocketServer.txt' using PigStorage('\t') as
     (fecha:chararray, uid:int, accion:chararray, tipo:int, marca:int);

userData = load 'proyecto/usersData.txt' using PigStorage('\t') as
         (uid:int, nombre:chararray, birthday:chararray, edad:int, sexo:chararray);

marcaNombreFilter = filter filtro by marcaCondicion;

marcaNombreUsuarios = foreach marcaNombreFilter generate uid;

marcaNombre = distinct marcaNombreUsuarios;

marcaNombreGroup = group marcaNombre All;

marcaNombreGeneral = foreach log generate marca, COUNT(uid) as totalUsuarios ;
```

Al final los indicadores obtenidos para cada marca son:

- Cantidad de usuarios que forman parte de su mercado objetivo.
- Cantidad de usuarios que interactuaron de alguna manera con su marca/producto.
- Cantidad de usuarios que pertenecen a su mercado objetivo y que realmente interactuaron con la marca/producto.
- Cantidad de usuarios que interactuaron con su marca/producto pero no

pertenecen a su mercado objetivo.

- Características más comunes de los usuarios que no pertenecen a su mercado objetivo pero que interactuaron con su marca/producto.

CAPÍTULO 4

Análisis

4.1 Decisiones de diseño

El formato para el log de actividades es resultado de los campos más comunes almacenados por sistemas de diferentes tipos que llevan este tipo de registro [11].Los datos de las empresas y los mercados objetivos de sus marcas o productos se obtienen de la base de datos de la aplicación y son previamente procesados para obtener el formato necesario que se definió en el capítulo de implementación y poder utilizarlo en el proceso de minería.

De la misma forma se pueden obtener los datos de usuarios, pero para el prsente proyecto se utilizó los datos disponibles en la red social de Facebook. Este proceso será reemplazado una vez que la aplicación de caso de estudio esté en producción, ya que los datos de usuarios estarán almacenados en la base de datos y se los puede obtener a través de un dump de la tabla.

Para el desarrollo de este trabajo hemos seleccionado Pig como la herramienta principal. Existe otra herramienta disponible llamada Hive, la cual provee herramientas para realizar sumarización de datos, consultas adhoc y análisis de grandes fuentes de datos almacenados en el sistema de archivos de Hadoop. Una de las ventajas de Hive es que su lenguaje QL es basado en SQL lo que facilita la adopción del mismo por parte de los usuarios, además permite a los programadores Map/Reduce utilizar sus procesos mapper y reducer específicos. Hive no promete disminuir la latencia en las consultas de datos.

Seleccionamos Pig en lugar de Hive para obtenre un mayor control en el orden en que se generan los porceso Map/Reduce ya que al ser PigLatin un lenguaje de flujo de procesos, el cambiar el orden en que se realizan los programas en Pig influye en los resultados de procesamiento final.

Durante la codificación se realizó todo dentro de un solo bloque Pig por cada marca, esta porción de código está embebida dentro de un programa en Java para obtener en una sola iteración todos los datos referentes a una marca y almacenarlos en una variable para al final del lazo unir todos los resultados en solo archivo que recopila la información de actividades del mes. Cada iteración resulta en varios procesos Map/Reduce que son administrados por la infraestructura de Pig. Es posible minimizar la cantidad procesos cambiando el orden de las sentencias PigLatin, ya que es un lenguaje de flujo de procesos. El orden utilizado para realizar la minería responde al orden que se seguiría haciendo el proceso manualmente.

4.2 Implementación

Este proyecto utiliza Hadoop en su versión 0.18 ya que es la más estable hasta el momento. La versión de Pig utilizada es 0.2. Durante el desarrollo del proyecto estuvo disponible versión 0.3 pero el proyecto se culminó con la versión anterior ya que igual permitía alcanzar los requerimientos establecidos. Es posible que Pig 0.3 contribuya a mejorar la eficiencia de este proyecto pero los resultados obtenidos con Pig 0.2 fueron satisfactorios.

4.3 Resultados

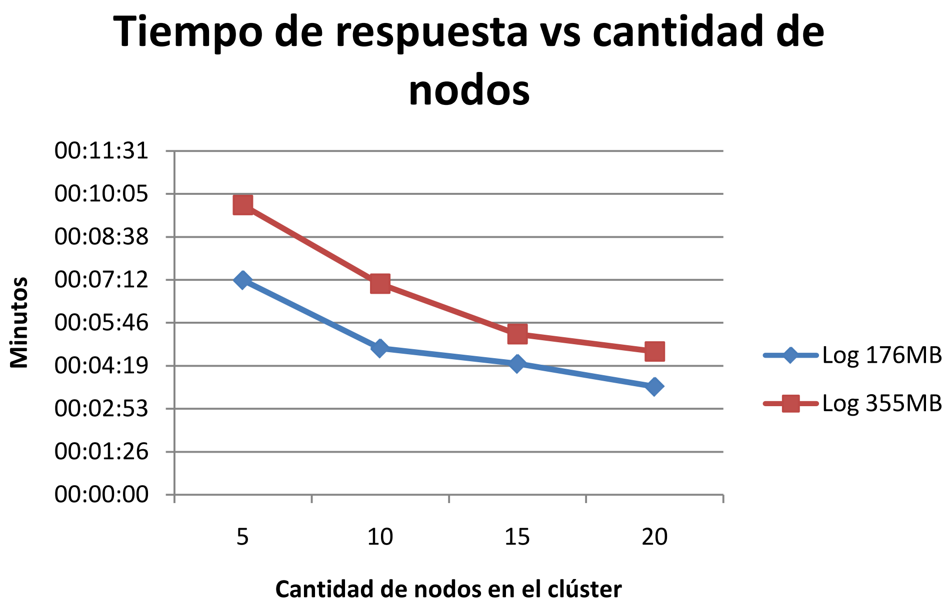
Se realizaron ejecuciones en varios escenarios descritos a continuación:

Las variables para cada escenario fueron las siguientes:

- Cantidad de nodos en el cluster
- Tiempo de respuesta en segundos

La Tabla 4 muestra los resultados obtenidos donde se analiza el tiempo de procesamiento para obtener resultados contra la cantidad de nodos levantados en el clúster.

Table 4. Resultados de ejecución



Estos resultados muestran el comportamiento con un log de 176MB de tamaño y otro de 355MB, los valores obtenidos pueden variar con tamaños de log mayores.

CONLUSIONES

1. Los indicadores obtenidos en los resultados permiten inferir el nivel de aceptación de un producto o marca dentro del entorno virtual que ofrece una aplicación social vinculada a los datos de una red social.
2. El beneficio que brinda este proyecto es optimizar el proceso de minería de logs a través de una plataforma distribuida con procesamiento en paralelo y el uso de Cloud Computing utilizando herramientas de código abierto disponibles en la web.
3. Este trabajo presenta la configuración más optima para obtener los resultados en el mejor tiempo y de manera más eficiente. Existen otros métodos para realizar este tipo de análisis pero el punto principal que ofrece este proyecto es utilizar las herramientas disponibles en la web y que son de fácil acceso.

RECOMENDACIONES

1. El resultado de este análisis es mucho más beneficioso si se tiene la mayor cantidad de características de usuario.
2. La minería de datos de archivos de logs usando la plataforma de Hadoop y Pig como herramienta nos ha sido útil para poder generar información para una aplicación en particular, sin embargo nuestro trabajo puede ser replicado para obtener otro tipo de información de carácter estadístico no para una empresa en particular pero quizás para investigaciones con valor social.
3. Para un trabajo futuro se podrían categorizar las marcas y productos para así obtener participación de mercado dentro de la aplicación, obtener más datos de usuario aprovechando los contenidos de las redes sociales.
4. El ignorar los usuarios que no colocan año de nacimiento puede introducir un sesgo en los datos de entrada. Es posible que ciertos grupos de usuarios tengan la tendencia a no colocar su año de nacimiento más que otros grupos de usuarios. Sería recomendable realizar un estudio (y posterior análisis de resultados) sobre le perfil de usuarios que no colocan su año de nacimiento para determinar si, al excluirlos, no se está ignorando uno o más grupos de usuarios

Bibliografía

[1]Shields Mike "In-Game Ads Could Reach $2 Bil.". Adweek. http://www.adweek.com/aw/national/article_display.jsp?vnu_content_id=1002343563 Abril, 2006.

[2] Ghemawat, S., Gobioff, H., y Leung, S. "The Google File System". En Memorias del 19th ACM Symposium on Operating Systems Principles. Lake George, NY-EE.UU., Octubre, 2003.

[3] C. Olston, B. Reed, U. Srivastava, R. Kumar and A. Tomkins. "Pig Latin: A Not-So-Foreign Language for Data Processing". ACM SIGMOD 2008 International Conference on Management of Data, Vancouver, Canada, June 2008.

[4] Dean, J. y Ghemawat, S. "MapReduce: Simplified Data Processing on Large Clusters". En memorias del Sixth Symposium on Operating System Design and Implementation (OSDI 2004), San Francisco, CA-EE.UU. Diciembre, 2004.

[5] http://aws.amazon.com

[6] http://aws.amazon.com/ec2/

[7] http://aws.amazon.com/s3/

[8] http://aws.amazon.com/s3/

[9] http://developers.facebook.com/

[10] http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=171

[11] Discovering Social Networks from Event Logs. Wil M.P. van der Aalst, Minseok Song