Minería de Logs de una Aplicación Multiusuario en Línea

Carlos Andrés Granda Bermúdez Luis Elicio Loaiza Pacheco Cristian Abad Robalino

cgranda@espol.edu.ec lloaiza@espol.edu.ec cabad@espol.edu.ec

Facultad de Ingeniería en Electricidad y Computación

Escuela Superior Politécnica del Litoral (ESPOL)

Campus Gustavo Galindo, Km 30.5 vía Perimetral, Guayaquil, Ecuador

Resumen

Este proyecto plantea realizar minería sobre los datos almacenados en logs o bitácoras de actividades de una aplicación que se encuentra vinculada a los datos de usuarios en una red social. La aplicación se desarrolla en un ambiente virtual y adapta el concepto de "In-Game Advertising", publicidad inmersa en el juego con el fin de tener un mayor protagonismo de las marcas e interactuar directamente con el cliente. El procesamiento se realiza de manera distribuida utilizando Hadoop como plataforma de desarrollo en conjunto con Pig y los servicios de Cloud Computing de Amazon. Como resultado final de este proceso se obtiene un reporte por cada marca o producto, el cual contiene indicadores básicos que permiten determinar el grado de aceptación de la marca o producto dentro del contexto en el que se desarrolla el juego. El análisis de resultados busca determinar la mejor configuración posible para obtener los reportes de la manera más eficiente y en el mejor tiempo posible. La configuración de trabajo propuesta presento resultados aceptables, aunque estos pueden ser mejorados.

Palabras Clave: Minería de Logs, Cloud Computing, Hadoop, Pig.

Abstract

This project proposes to perform data mining on activity logs of an application that is linked to the data of users in a social network. The application is developed in a virtual environment and adapts the concept of "In-Game Advertising", Advertising immersed in the game in order to have a higher profile of brands and interact directly with the customer. The processing is done in a distributed way using Hadoop as a map-reduce processing platform in conjunction with Pig and cloud computing services from Amazon. The end result of this process, is a report for each brand or product, which contains basic indicators for determining the acceptability of the brand or product within the context in which the game develops. The result analysis seeks to determine the best configuration to get the reports in the most efficient and the best time possible. The structure proposed presented acceptable results, although they can be improved.

1. Introducción

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.

Actualmente en la Web ha habido un auge de aplicaciones sociales multiusuario en tiempo real. Tanto las redes sociales como los juegos comparten información valiosa sobre las relaciones entre usuarios. La Web 2.0 ha permitido a su vez que los usuarios tengan protagonismo y sean parte fundamental de las aplicaciones. Los usuarios al interactuar con las aplicaciones generan eventos y realizan actividades que demuestran su comportamiento.

Hemos tomado la iniciativa de crear una aplicación web, para que clientes y potenciales clientes se involucren de manera más rápida y efectiva con marcas y productos de empresas alrededor del mundo. Esto crea una manera diferente de publicidad donde las personas en general participan e interactúan con objetos relacionados a marcas.

Una de las partes medulares del proyecto es retroalimentar a las Empresas (marcas) con indicadores de interés específicos. Estos indicadores se obtienen a través de la minería de "logs" (bitácora) de actividades de los usuarios de la aplicación. El análisis de los datos dada por los *logs* permite saber características de las personas y sus preferencias.

Estos logs de actividades almacenan a diario gran cantidad de información alcanzando tamaños de hasta miles de MegaBytes (MB), por lo que es necesario procesarlos de la manera más eficiente. Hadoop[1] es una herramienta para procesar grandes cantidades de datos de manera paralela y optimizando recursos.

2. Diseño de la solución

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.

2.1 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.

2.2 Plataforma de Cloud Computing

Otro concepto clave para realizar este proyecto “Cloud Computing” y la herramienta para este propósito son los Amazon Web Services [5] , específicamente el Elastic Computing Cloud (EC2) [6] para levantar un cluster Hadoop y el Simple Storage Service (S3) [7] 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.

2.3 Flujo de procesos

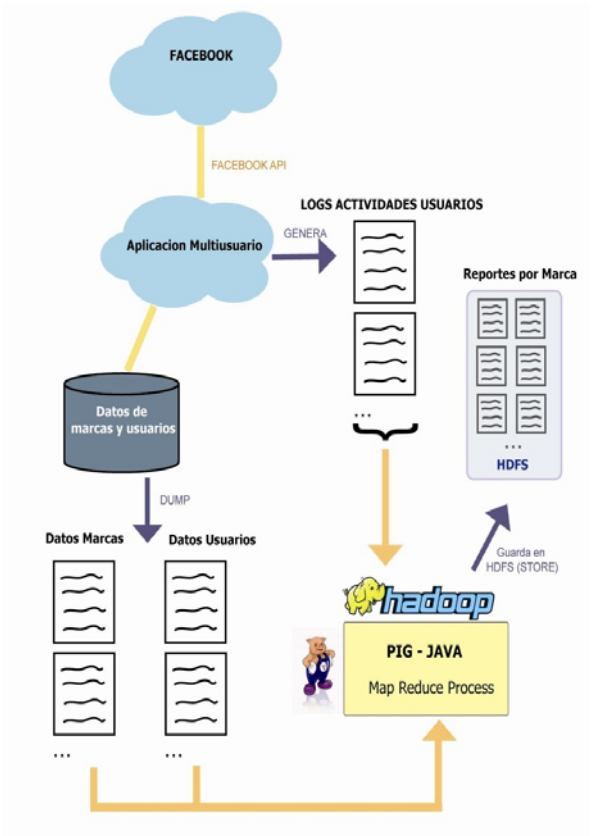


Figura 1. Flujo de procesos

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[8], 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.

3. Implementación

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	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 M	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
12-04-2009 12:30:34	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 [9] 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.

4. Conclusiones y Resultados

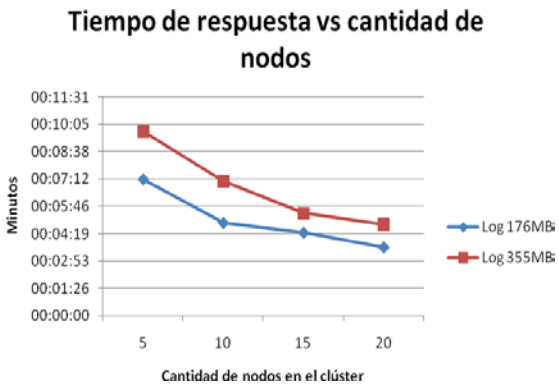
Se realizaron ejecuciones en varios escenarios descritos a continuación:

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.

Tabla 4. Resultados de ejecución



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 óptima 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.

5. Recomendaciones

- El resultado de este análisis es mucho más beneficioso si se tiene la mayor cantidad de características de usuario.
- 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.
- 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.
- 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 el 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

6. Referencias

[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] Servicios de computación en la nube de Amazon disponibles en http://aws.amazon.com.

[6] Servicio de computación en la nube, "Elastic Computing Cloud" disponible en http://aws.amazon.com/ec2/ .

[7] Servicio de computación en la nube, "Simple Storage Service" disponible en http://aws.amazon.com/s3/

[8] Guía para desarrolladores de la plataforma social de Facebook disponible en http://developers.facebook.com/ .

[9] Imágenes de máquinas (máquinasvirtuales) de Amazon disponibles para desarrolladores en http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=171

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