Informe de Materia de Graduación

“Procesamiento Masivo y Escalable de Datos”

Presentado por:

- Luis Loaiza
- Carlos Andrés Granda

“MINERÍA DE LOGS DE UNA APLICACIÓN SOCIAL MULTIUSUARIO EN LÍNEA”

Introducción

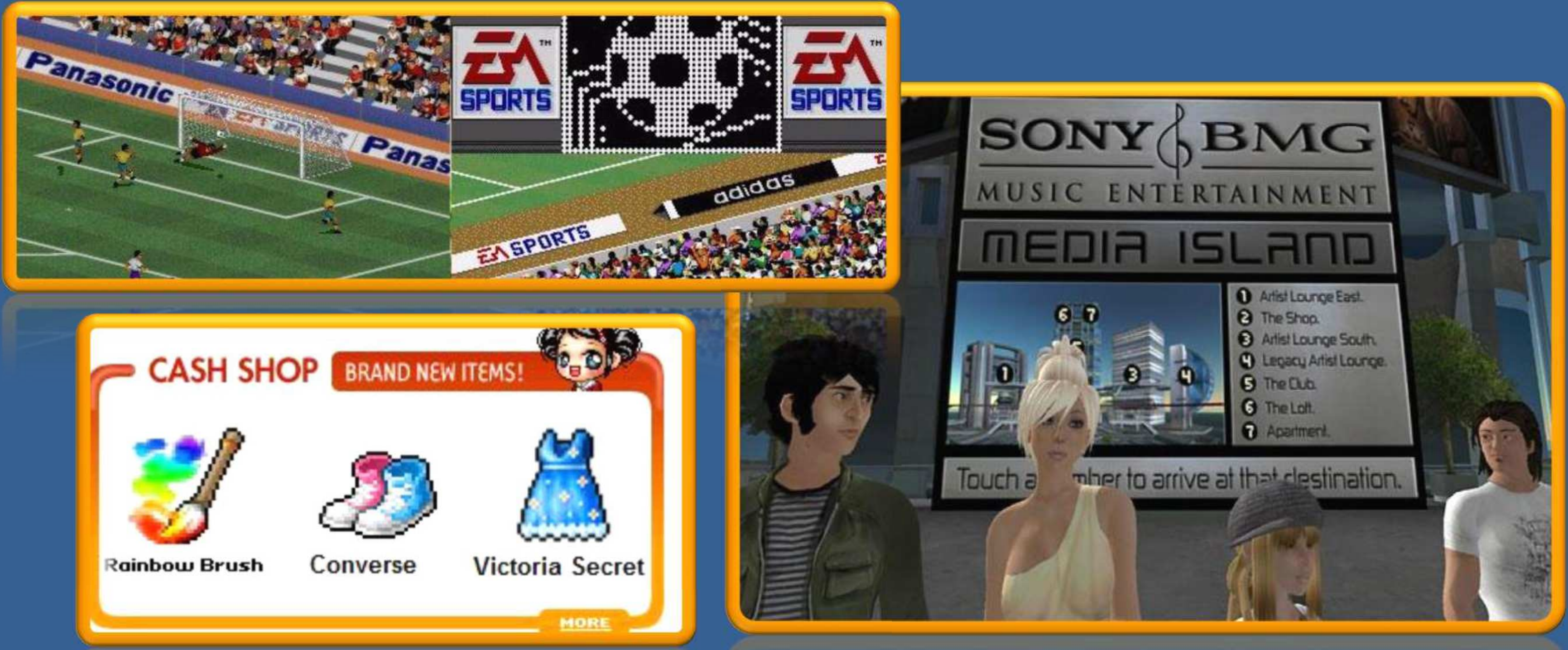
- **Minería** es el proceso de extraer información que se encuentra implícita en los datos.
- **Aplicación social**: son de uso masivo y están vinculadas a las redes sociales.
- **Logs o bitácoras** son archivos que almacenan información con respecto a actividades siguiendo un formato específico.

Objetivo

- Realizar minería sobre los archivos de log de una aplicación social que utiliza el concepto de In-Game Advertising, para obtener indicadores de aceptación de marcas o productos, que se encuentran representados de manera virtual dentro de la aplicación.

In-Game Advertising

- Estrategia de Marketing que incluye publicidad sutilmente dentro del contexto de un juego o aplicación social.



Aplicaciones similares a la del caso de estudio.

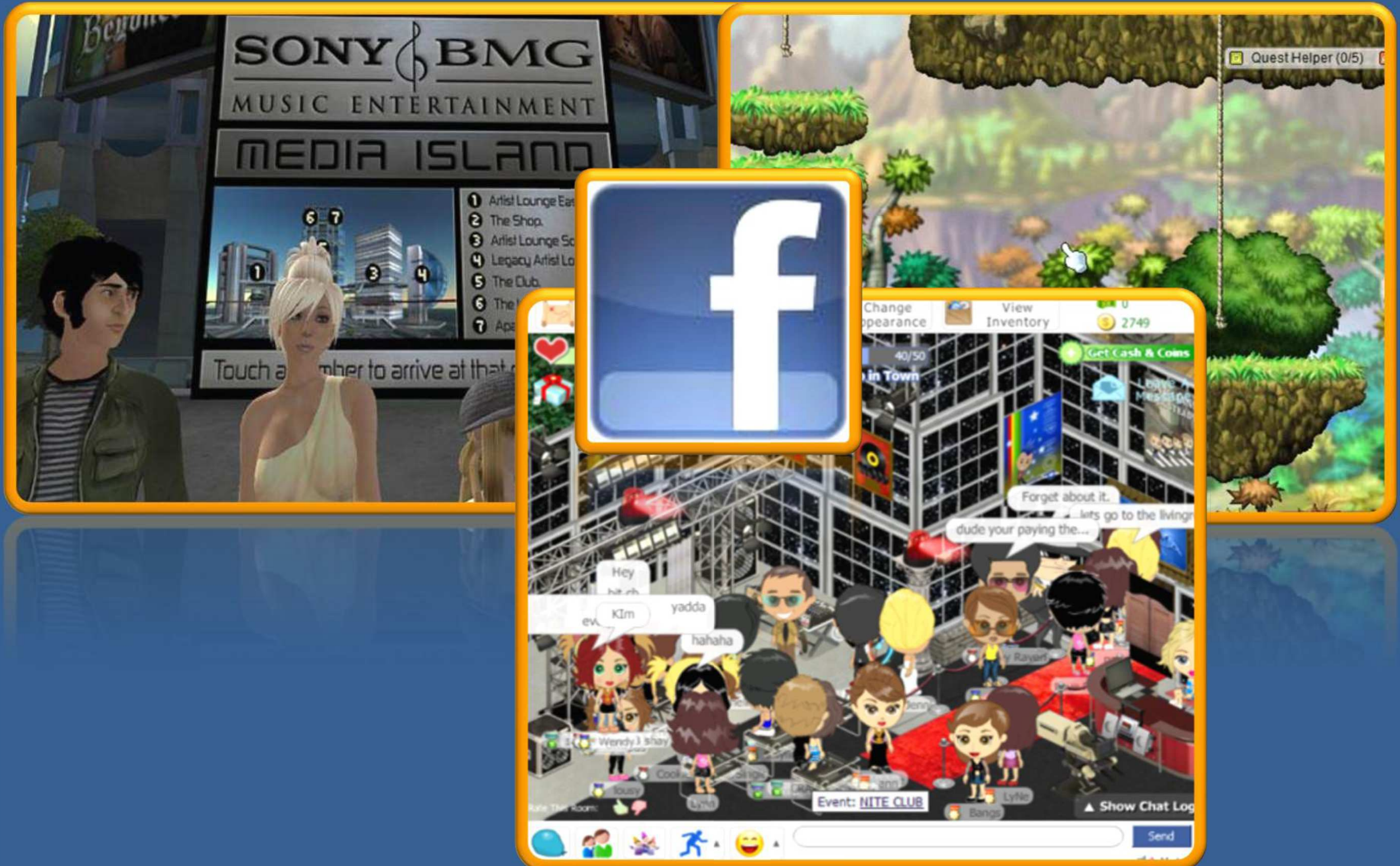
FarmVille de ZINGA

- Usuarios diarios
  13’ 592.404
- Usuarios mensuales
  37’ 659 .165
- Crecimiento Diario 2.12

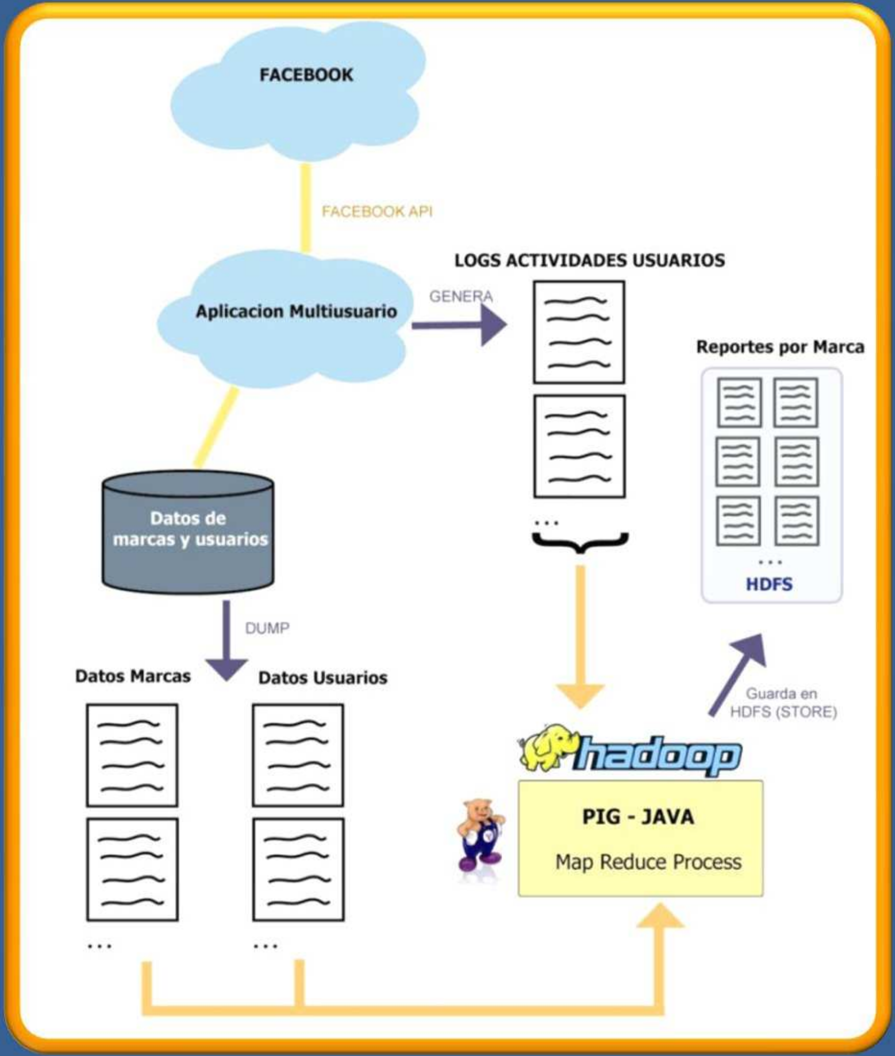
Pet Society de PlayFish

- Usuarios diarios
  4’543.851
- Usuarios mensuales
  16’634.509
- Crecimiento Diario 0.78

Características de la aplicación de caso de estudio.



Flujo de Procesos



1. Generar archivo de usuarios
2. Generar archivo de empresas
3. Obtener log de actividades
4. Subir los archivos al S3
5. Cargar el proceso en EC2
6. Generar Reportes

Recursos Utilizados

- Facebook API
- Amazon Web Services:
  - Elastic Cloud Computing (EC2)
  - Simple Storage Server (S3).
- Apache Hadoop
- Apache Pig

Datasets de entrada

Marca o Producto por Empresa

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

Usuario

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

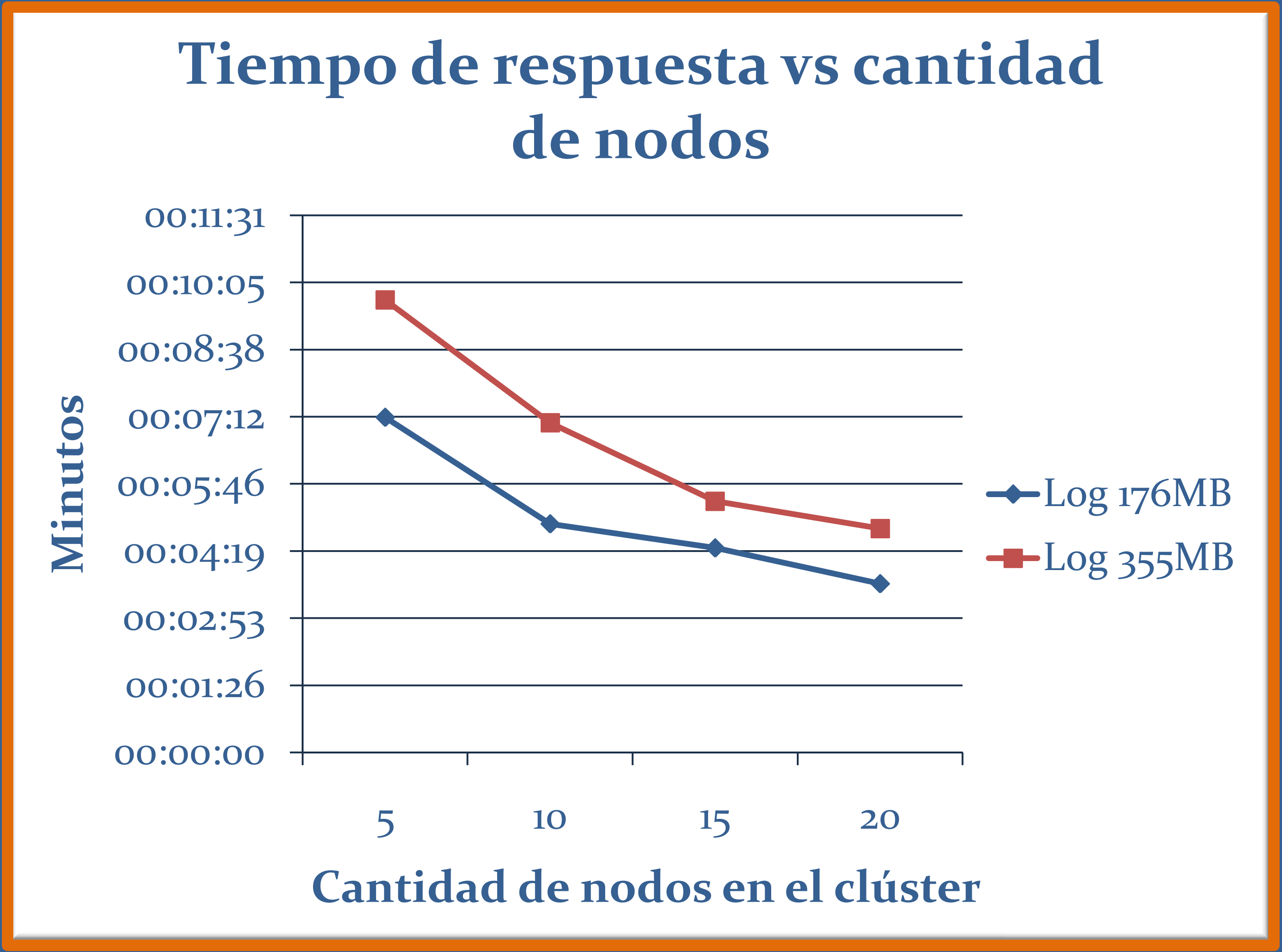
Log de Actividades

DateTime	IdUsuario	Actividad	Tipo	Marca
12	126	Personalizar	1	12

Código Pig

```
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;
```

Resultados


Tiempo de respuesta vs cantidad de nodos
Minutos
00:11:31
00:10:05
00:08:38
00:07:12
00:05:46
00:04:19
00:02:53
00:01:26
00:00:00
5
10
15
20
Cantidad de nodos en el clúster
Log 176MB
Log 355MB

Conclusiones

- Los indicadores obtenidos permiten inferir nivel de aceptación.
- Optimiza el proceso de minería de logs.

Recomendaciones

- Mayor cantidad de características de usuario.
- Este trabajo puede ser replicado para obtener otro tipo de información estadística.
- Para trabajo futuro se podrían categorizar las marcas y productos.