RedES: Informes

Siguiendo el desarrollo de la red social se procederá con la fase de informes. Esta fase será desarrollada íntegramente en *python*.

Hasta el momento la gestión de *logs* se ha realizado mediante ficheros en texto plano. Se desea crear un conjunto de consultas *mapreduce* que permita obtener informes con información especifica de los *logs* obtenidos hasta el momento. Para tal fin, se han almacenado los *logs* en un clúster de tipo Hadoop que permite realizar consultas de forma eficiente.

Se comenzará implementando las consultas necesarias para el análisis de los *logs* de Apache. Se provee un archivo de muestra de los *logs* almacenados por apache. Cada línea se corresponde con una petición al servidor Apache donde cada uno de los datos almacenados están separados por puntos y comas, ";".

Requisitos de la práctica (Consultas)

Se entregará dos archivos por cada consulta, uno para el *mapper* y otro para el *reducer*. Los nombres de los archivos deberán ser mX.py para el *mapper* y rX.py para el *reducer* donde X se corresponde con el número de consulta.

El archivo debe poder ser ejecutable desde el terminal, por lo que deberá comenzar con la línea *#!/usr/bin/python* (o la ruta correspondiente al ejecutable de *python*).

El fichero apache.log contiene el repositorio de información semi-estructurada sobre la que se realizarán las consultas. Se simulará realizar las consultas sobre el clúster a través del archivo apache.log. En ningún caso será necesario realizar las consultas sobre *Hadoop*. Para simular el mapreduce se utilizará el siguiente comando en un sistema con kernel de Linux.

```
cat testfile.txt | ./mapper.py | sort | ./reducer.py
```

En Windows existen similares a cat y sort que se pueden utilizar.

La librería *sys* de Python es necesaria para acceder a la información que se pasa por stdin tanto al *mapper* como al *reducer*: *sys.stdin*

Se ha de tener en cuenta que es posible que existan errores en los datos proporcionados. En caso de encontrar un error se deberá:

- Más o menos columnas: Se salta la fila.
- Cadenas de texto donde se espera valores numéricos: Se salta la fila.
- Espacios en blanco al principio y al final de una cadena: Cadenas de texto con mismo contenido con o sin espacios antes y/o después de la cadena deben ser equivalentes.

Consultas

1. Número de peticiones de tipo GET
2. Número de peticiones por tipo (GET, POST, PUT, etc.)
3. Media/porcentaje de peticiones por tipo (GET, POST, PUT, etc.)

4. Tipos de respuesta por familia (2XX, 3XX, 4XX y 5XX) por tipo de petición (GET, POST, PUT, etc.)
5. Tipos de archivo (.gif, .html, etc.) no encontrados (404) en un hora determinada.
6. Incremento/decremento por hora (en porcentaje) del tráfico en bytes (último dato).
7. Top tres dominios que más tráfico tiene en bytes (último dato).
8. Top tres tipos de archivos (.gif, .html, etc.) que más tráfico produce en bytes (último dato) por dominio.
9. Balance de tráfico en bytes (último dato) por dominio. Descargas (GET) menos subidas (POST).

Normativa de realización, entrega y evaluación de la práctica:

- La práctica se realizará y entregará en grupos de hasta dos integrantes.
- La práctica se realizará en python y haciendo uso de un terminal bash.
- La entrega se compondrá de un único fichero ZIP con cada uno de los mappers y reducers de cada consulta.
- El directorio y fichero zip será renombrado con el número de la práctica P#, seguido del número del grupo G#, y finalmente seguido con el nombre y el primer apellido de los alumnos integrantes del grupo, separados mediante guiones bajos '_'.

  Ejemplo: *P4_G25_Quijote_de_la_Mancha-Sancho_Panza.zip*

- Se considerará suspensa toda práctica cuyo fichero comprimido no contenga los ficheros fuente.
- Las prácticas entregadas fuera de plazo, serán calificadas sobre 9. Por cada día de retraso en la entrega se reducirá el rango de calificación en 0,2 puntos.
- Cualquier sospecha de COPIA entre dos o más prácticas o de código obtenido en internet derivará en la calificación de 0 para todos los alumnos involucrados en la evaluación en curso y la siguiente. En caso de que el alumno tenga duda de que el código pueda ser susceptible de ser entendido como copia, consultar con el profesor antes de la entrega.