# Web Server logs Analysis

De forma general, un server log es un archivo de log generado por el servidor con una lista de las actividades que se ejecutan. En este caso tenemos un web server log el cuál mantiene un historial de las peticiones realizadas a la página. Este tipo de server logs tienen un formato standard (Common Log Format). Y es una práctica general, el analizar estos logs para sacar distintas conclusiones, localizar ataques, errores comunes, etc.

En nuestro caso tenemos el dataset de los web server logs de la NASA. Qué están compuestos por este tipo de registros:

133.43.96.45 - - [01/Aug/1995:00:00:23 -0400] "GET /images/launch-logo.gif HTTP/1.0" 200 1713

Por lo que tenemos los siguientes campos:

1. **Host:** 133.43.96.45
2. **User-identifier:** en este dataset, todos estos campos estarán con un "-" que significa que faltan esos datos, por lo que obviaremos este campo.
3. **Userid:** al igual que el anterior campo, también será obviado.
4. **Date:** 01/Aug/1995:00:00:23 -0400, como podemos ver está en formato dd/MMM/yyyy:HH:mm:ss y el campo final "-0400" sería el timezone que en este caso omitiremos, además haremos una transformación de los meses a forma numérica.
5. **Request Method:** GET, existen distintos métodos de petición aquí puedes obtener más información: link
6. **Resource:** /images/launch-logo.gif, sería el recurso al que se accede en esta petición.
7. **Protocol:** HTTP/1.0, y por ultimo en esta parte entre comillas tendríamos el protocolo utilizado al ser logs de 1995, seguramente sea el único protocolo utilizado.
8. **HTTP status code:** 200, existen distintos códigos de estado de HTTP en el link a continuación tienes más información: link
9. **Size:** 1713, y como ultimo campo tendríamos el tamaño del objeto recibido por el cliente en bytes. En casos de error del cliente, este campo no se encuentra por lo que al igual que en los userid, será indicado con un "-", tenerlo en cuenta.

Ahora que ya entendemos que se encuentra dentro de nuestro web server log, vamos a pasar a analizarlo. Primero debemos cargar el archivo como un archivo de texto normal y realizar las transformaciones pertinentes, a la hora de limpiar y estructurar nuestro dataset utilizaremos expresiones regulares para recoger los campos que necesitamos.

Guardaremos nuestro nuevo DataFrame ya estructurado en formato parquet. Y de este leeremos para realizar nuestro análisis.

Consultas a realizar:

- ¿Cuáles son los distintos protocolos web utilizados? Agrúpalos.
- ¿Cuáles son los códigos de estado más comunes en la web? Agrúpalos y ordénalos para ver cuál es el más común.
- ¿Y los métodos de petición (verbos) más utilizados?
- ¿Qué recurso tuvo la mayor transferencia de bytes de la página web?
- Además, queremos saber que recurso de nuestra web es el que más tráfico recibe. Es decir, el recurso con más registros en nuestro log.
- ¿Qué días la web recibió más tráfico?
- ¿Cuáles son los hosts son los más frecuentes?
- ¿A qué horas se produce el mayor número de tráfico en la web?
- ¿Cuál es el número de errores 404 que ha habido cada día?

**Links de ayuda e interés**

https://regex101.com/ - Para trabajar con expresiones regulares, muy útil.

Ejemplos de ejercicios de análisis de web server logs:

- Ejemplo Python 1
- Ejemplo Python 2
- Ejemplo Python 3
- Ejemplo Scala