Captura y obtención de datos
Métodos de recopilación

Web Server Logging

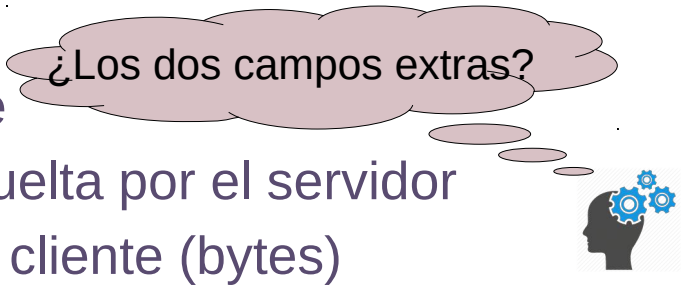
- Método tradicional de recolectar datos de uso en la web
  - Los servidores registran algunas de sus transacciones en un archivo de texto (archivo *log*)
    - Actividades
    - Encabezados HTTP
  - Accesos desde navegador o robots
- Formato de los archivos
  - Formatos libres (NCSA, Extended Format)
  - Formato propietario

Web Server Logging

- ¿Cómo funciona?
  - Recolección de datos
    - Se registra cada petición realizada al servidor
  - Procesamiento de datos
    - Los datos se almacenan en bruto, sin estructura
    - Procesarlos a un formato estructurado y legible
  - Análisis
    - Aplicar técnicas de análisis para comprender como interactúan los usuarios con el sitio web
  - Informe
    - Visualización de la información obtenida

Web Server Logging

- Formato de los archivos
  - Common Log Format (NCSA)
    - Cada línea:
      - Host → dirección IP del host que hace la petición
      - Ident → identificador
      - Authuser → useId de la persona que hace la petición
      - Date → fecha, hora y zona horararia
      - Request → línea de petición del cliente
      - Status → código del estado HTTP devuelta por el servidor
      - Bytes → tamaño del objeto devuelto al cliente (bytes)




```
84.245.59.290 – – [01/Oct/2018:08:39:04 +0200] “GET
/module/CLNEWMSG/css/bubble.css?1251290622 HTTP/1.1” 304 136
“https://www.axarnet.es/alojamiento-web-linux/” “Mozilla/5.0
(Windows NT 6.1; rv:24.0) Gecko/20100101 Firefox/24.0”.
```

Web Server Logging

- Formato de los archivos
  - Extended Log File Format (WC3)
    - Más información y flexibilidad que el NCSA
    - Secuencia de líneas con caracteres ASCII terminados en secuencias LF o CRLF
    - Cada línea es una directiva o una entrada
    - La directivas aparecen marcadas con #
      - Registran información sobre el proceso de registro en sí
    - Las entradas son secuencias de campos relacionados con una transacción HTTP
      - Separados por espacios en blanco
      - Si un campo no se usa, se pone el guión (-)

Web Server Logging

- Formato de los archivos (W3C)
  - Directivas
    - Version: *<integer>.<integer>*
      - La versión del formato de archivo utilizado
    - Fields: *[<specifier>...]*
      - Especifica los campos que se van a registrar en cada entrada
    - Software: *string*
      - Identifica el software que genera el log
    - Start-Date: *<date> <time>*
      - La fecha y hora a la que se empezó el log
    - End-Date:*<date> <time>*
      - La fecha y hora en la que se terminó el log

Web Server Logging

- Formato de los archivos (W3C)
  - Directivas
    - Date:<*date*> <*time*>
      - La hora y fecha en la que se añadió la entrada
    - Remark: <text>
      - Comentarios. La herramientas de análisis la ignoran
  - Las directivas Version y Field deben aparecer siempre y antes de cualquier entrada.

Web Server Logging

- Formato de los archivos (W3C)
  - Forma de las entradas de la directiva #Fields
    - Identificador | prefijo-identificador
  - Prefijos posibles en #Fields
    - c → cliente
    - s → servidor
    - r → remoto
    - cs → de cliente a servidor
    - sc → de servidor a cliente
    - sr → de servidor a servidor remoto (usado por proxis)
    - rs → de servidor remoto a servidor (usado por proxis)
    - x → identificador específico de una aplicación

Web Server Logging

- Formato de los archivos (W3C)
  - Identificadores posibles en #Fields sin prefijo
    - date → fecha de fin de la transacción
    - time → hora de fin de la transacción
    - time-taken →tiempo que tarda la transacción en completarse
    - Bytes → nº de bytes transferidos
    - Cached → 0 indica un fallo de cache
    -

Web Server Logging

- Formato de los archivos (W3C)
  - Identificadores posibles en #Fields con prefijo
    - ip → dirección IP y puerto
    - dns → name del DNS
    - status → código de estado
    - comment → comentario devuelto con código de estado
    - method → método
    - uri → URI
    - uri-stem → Solo parte de la raíz del URI (se omite la consulta),
    - uri-query → Solo parte de la consulta del URI

Web Server Logging

- Formato de los archivos (W3C)
  - Ejemplo

```
#Version: 1.0
#Date: 12-Jan-1996 00:00:00
#Fields: time cs-method cs-uri
00:34:23 GET /foo/bar.html
12:21:16 GET /foo/bar.html
12:45:52 GET /foo/bar.html
12:57:34 GET /foo/bar.html
```

Web Server Logging

- Formato de los archivos
  - IIS
    - Campos
      - Client IP address → Dirección IP del cliente
      - User name → Nombre del usuario que accede al servidor.
      - Date → Fecha en la que ocurre la actividad
      - Time → Hora local en la que ocurre la actividad
      - Service and instance → Nombre del servicio de Internet y el número de instancia que se estaba ejecutando en el cliente.
      - Server name → Nombre del servidor en el que se generó la entrada del archivo de registro.
      - Server IP address →Dirección IP del servidor en el que se generó la entrada del archivo de registro.
      - Time taken → Tiempo, en ms, que duró la acción

Web Server Logging

- Formato de los archivos
  - IIS
    - Campos
      - Client bytes sent → N.º de bytes enviados por el cliente
      - Server bytes sent → Nº de bytes enviados por el servidor
      - Service status code → Un valor de 200 indica que la petición tuvo éxito
      - Windows status code → Un valor de 0 indica que la petición tuvo éxito
      - Request type → El tipo de petición (verbo)
      - Target of operation → El destino de la operación
      - Parameters → Parámetros que se pasan a un script

Web Server Logging

- Formato de los archivos (IIS)

```
192.168.114.201, -, 03/20/05, 7:55:20, W3SVC2, SERVER,
172.21.13.45, 4502, 163, 3223, 200, 0, GET, /DeptLogo.gif, -,
```

Web Server Logging

- Ventajas
  - Almacenan la información por defecto
    - No es necesario modificar la web
  - Datos en el propio servidor de la empresa
    - Formato estándar en vez de propietario
    - Datos de nuestra propiedad
  - Contienen información sobre las visitas de las arañas de los buscadores
  - No requieren búsquedas de DNS adicionales
    - No hay llamada a servidores externos que ralenticen los tiempos de carga

Web Server Logging

- Ventajas
  - El servidor registra de manera confiable cada transacción
    - No depende de la cooperación de los navegadores de los usuarios
  - Acceso a datos sin procesar
    - Mayor flexibilidad en el análisis y personalización de métricas
  - Mejor cumplimiento de la privacidad (RGPD)
    - No se basa en cookies ni en seguimiento

Web Server Logging

- Contras
  - El calculo de visitas se hace por IP
    - Pueden ser más o menos que las reales
      - Se cuentan los accesos de los robots
      - No se cuentan los accesos desde cache
  - Complejidad técnica
    - Experiencia para procesar y analizar los datos de manera efectiva
  - Contexto limitado
    - No se capturan cierto tipo de interacciónes
  - Retención de datos
    - Gran volumen de datos difícil de manejar

Web Server Logging

- Herramientas para extraer y analizar registros
  - Deep Log Analyzer http://www.deep-software.com
  - SawMill (www.sawmill.net)
  - AWSstats (http://awstats.sourceforge.net)
  - Webalizer

Page tagging

- Programa en el lado del cliente que se encarga de recopilar la información
  - Comandos incrustado (*embedded scripts*)
  - *Add-ons* y *plugins* para el navegador
- Formato más utilizado
  - Incluir un fragmento de código JavaScript en el código HTML de la página
    - Rastrea la actividad del usuario y la almacena en una *cookie*
  - La información se envía a un servidor de procesamiento

Page tagging

- Ejemplo de código para utilizar con


```
<script async
src="https://www.googletagmanager.com/gtag/js?id=TAG_ID"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());

  gtag('config', 'TAG_ID');
</script>
```

Page tagging

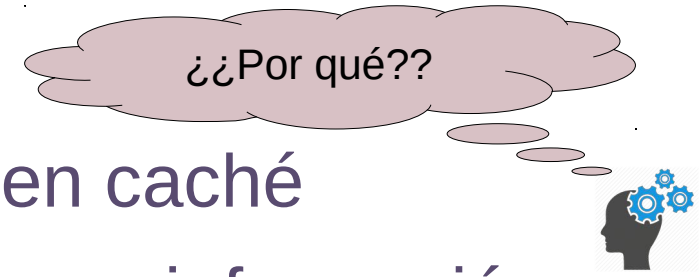
- Como funciona
  - El código JavaScript se ejecuta cada vez que se carga una página web
    - Independientemente de donde provenga el acceso
  - El script devuelve la llamada al servidor del sitio web y la pasa la información del usuario
    - Llamadas asíncronas que no afectan al rendimiento
  - El código asigna una cookie persistente para almacenar la identificación del cliente
    - Cuando el usuario vuelve a visitar la página, la información almacenada en la cookie lo identifica

Page tagging

- Como funciona
  - Si los datos se recopilan a través de un tercero, como Google Analytics, los datos se transfieren al servidor de ese tercero.

Page tagging

- Ventajas



  - No influye el que una página esté en caché
  - Los scripts pueden tener acceso a información adicional
    - Tamaño de pantalla, precio de los bienes comprados, …
  - Puede informar sobre eventos que no implican una solicitud
    - Interacción dentro de películas Flash
    - Finalización parcial de un formulario
    - Eventos de ratón (onClick, onMouseOver, onFocus,...)

Page tagging

- Ventajas
  - Gestiona el proceso de asignación de cookies a los visitantes
    - Los archivos de log hay que configurar el servidor para ello
  - Está disponible para empresas que no tienen acceso a su propio servidor
  - Servicio brindado a través de software como servicio (SaaS)
    - No necesita mantenimiento
    - Preferido en los sitios web pequeños y medianos

Page tagging

- Ventajas
  - La mayoría de los dispositivos aceptan JavaScript
  - Pocos usuarios tienen desactivado JavaScript
  - Las herramientas más populares utilizan este sistema
    - Google Analytics
- Contras
  - Todas las páginas deben incluir el código
  - Se pueden eliminar las cookies

Cookies

- Archivo que se descarga en el ordenador del usuario que visualiza una página web
  - Se ocupa de almacenar y recolectar datos
    - Patrones y hábitos de navegación de los usuarios
    - Datos sobre el ordenador, navegador, tipo de usuario
- Las herramientas basadas en tags de JavaScript las utilizan

Cookies

- Tipos de cookies
  - Según el emisor
    - Cookies propias
      - Las gestiona y envía el mismo sitio web
    - Cookies a terceros
      - Las envía una entidad diferentes (Google, Youtube)
  - Según la duración
    - Cookies de Sesión
      - Se borran cuando el usuario cierra el navegador
    - Cookies persistentes
      - Tienen un periodo de tiempo específico
      - Se pueden borrar

Cookies

- Tipos de cookies
  - Según la finalidad
    - Cookies técnicas
      - Aquellas que permiten controlar el tráfico y la comunicación de datos
      - Permiten saber los sitios web de dónde viene el usuario
    - Cookies de personalización
      - Filtran a los usuarios según sus características recogidas
    - Cookies de análisis
      - Permiten cuantificar el n.º de usuarios, analizar el tráfico del sitio web, hacer un seguimiento del sitio web
    - Cookies publicitarias
      - Gestionan los espacios publicitarios que el emisor ha incluido gracias a la información del comportamiento de los usuarios

Cookies

- Tipos de cookies
  - Otros
    - Cookies seguras
      - Se utilizan en conexiones HTTPS
      - Acumulan información cifrada para evitar que los datos sean vulnerables
    - Cookies zombie
      - Se guardan en el dispositivo y no en el navegador
      - Almacena todo lo que hace el usuario en el dispositivo
      - Amenazan la privacidad y seguridad de los usuarios

Cookies

- Ventajas
  - Aceptadas por la mayoría de los dispositivos
  - Permite mantener un histórico de visitas del usuario
- Contras
  - El borrado de las cookies por parte del usuarios
  - Algunos firewalls o programas de seguridad bloquean las cookies.

Logs de ISP

- ISP
  - Provedores de servicios de Internet
  - Empresas que proporcionan acceso a internet
- Sus logs almacenan información de
  - Actividad en la red
  - Uso de de los servicios por los clientes
- Ejemplo
  - Registro de conexiones
    - Conexiones de red de los clientes
    - Dirección IP, fecha, duración, tipo conexión

Logs de ISP

- Ejemplos
  - Registros de tráfico
    - Tráfico generado por los clientes
      - Direcciones IP origen/destino, puertos, protocolo red …
  - Registro direcciones IP
    - Direcciones IP asignadas a los clientes
  - Registro de errores y eventos
    - Eventos relacionados con la infraestructura de red
      - Fallos de red, cortes de servicio, problemas, …
  - Registro autenticación
    - Intentos de autenticación de los clientes
      - Nombres de usuario, contraseñas, fechas, resultados

Logs de ISP

- Ventajas
  - Permiten seguir la actividad de los usuarios fuera del sitio web
- Contras
  - No se pueden conocer los perfiles demográficos de los usuarios que visitan un sitio web
    - Se trabaja con muestreos

Panel de usuarios

- Como funciona
  - Se instala un software de recogida de datos en los sistemas de algunos usuarios
    - Recoger estadísticas sobre la audiencia del sitio web
  - Se utiliza como muestreo para determinar el número de usuarios que visitan el sitio web en un periodo determinado

Panel de usuarios

- Ventajas
  - Proporciona información demográfica usuarios
- Desventajas
  - Datos no precisos
  - Los resultados no pueden darse en tiempo real

Barra del navegador

- Los datos se obtienen a partir de las extensiones que se instalan los usuarios en la barra de navegación de su navegador
  - Capturan todo lo que ocurre en el navegador
- Ventajas
  - Permite conocer datos de la competencia de forma gratuita
- Contras
  - Datos inexactos → perfiles geek
  - Segmentos hiper representados

Resumen

- Diversas forma de capturar los datos
  - La más habituales
    - Analizar la información almacenada en los ficheros *log* de los servidores
    - Introducir algún tipo de código en el sitio web que se encargue de recopilar la información
  - Cada una unos resultados diferentes
    - Elegir en función del sitio y los objetivos