MMM006-340161 Web Analytics
Programme: Data Engineering

JACOBS
UNIVERSITY

Fall Semester 2018

Lecture 2: Data Collection Techniques

Instructor: Dr. Bendick Mahleko

b.mahleko@jacobs-university.de

Objectives

Objectives:

- Describe data collection techniques used in Web Analytics
- Evaluate the data collection techniques

Log file analysis

NCSA Common log file format

The common log file format records the following information

- ip address of the computer that requested information
- authenticated user
- date and time at which the transaction was completed
- time taken for the transaction
- request – the request line
- bytes transferred
- status

Log file analysis

W3C extended log file format

This log file format is customizable

Some information collected by the extended log file may include

- ip address of the computer that requested information
- date and time at which the transaction was completed
- time taken for the transaction
- bytes transferred
- flag to indicate if cache was used or not
- referrer

Log file analysis

IIS Log format

The IIS log format records the following information:

- ip address of user
- user name
- date
- time
- service
- computer name
- ip address of server
- time taken in ms
- bytes received
- bytes sent
- status
- request type
- target of operation

Example

Log files continued ...

%a - Remote IP address

%A - Local IP address

%b - Bytes sent, excluding HTTP headers, or '-' if zero

%B - Bytes sent, excluding HTTP headers

%h - Remote host name (or IP address if enableLookups for the connector is false)

%H - Request protocol

%l - Remote logical username from identd (always returns '-')

%m - Request method (GET, POST, etc.)

%p - Local port on which this request was received. See also %{xxx}p below.

%q - Query string (prepended with a '?' if it exists)

%r - First line of the request (method and request URI)

%s - HTTP status code of the response

%S - User session ID

%t - Date and time, in Common Log Format

%u - Remote user that was authenticated (if any), else '-'

%U - Requested URL path

%v - Local server name

%D - Time taken to process the request, in millis

%T - Time taken to process the request, in seconds

%F - Time taken to commit the response, in millis

%I - Current request thread name (can compare later with stacktraces)

Source: Tomcat documentation: https://tomcat.apache.org/tomcat-7.0-doc/config/valve.html

Pros and cons of log file analysis

Pros:

- Data ownership – website owner has full control privacy of information
- Logs available backwards – allows historical analysis to be performed
- Saves web crawler behavior

Cons:

- If data is retrieved from cache, visit is not recorded
- Log files can grow to be very large over time

Page tagging

Insert a small JavaScript code to all pages that are to be tracked.
When the page is loaded, the following data is sent to the database

- page id
- timestamp when page was loaded
- user origin (search engine, referral link)
- ip address
- Technical details
  - browser info
  - operating system info
  - screen resolution
  - colors

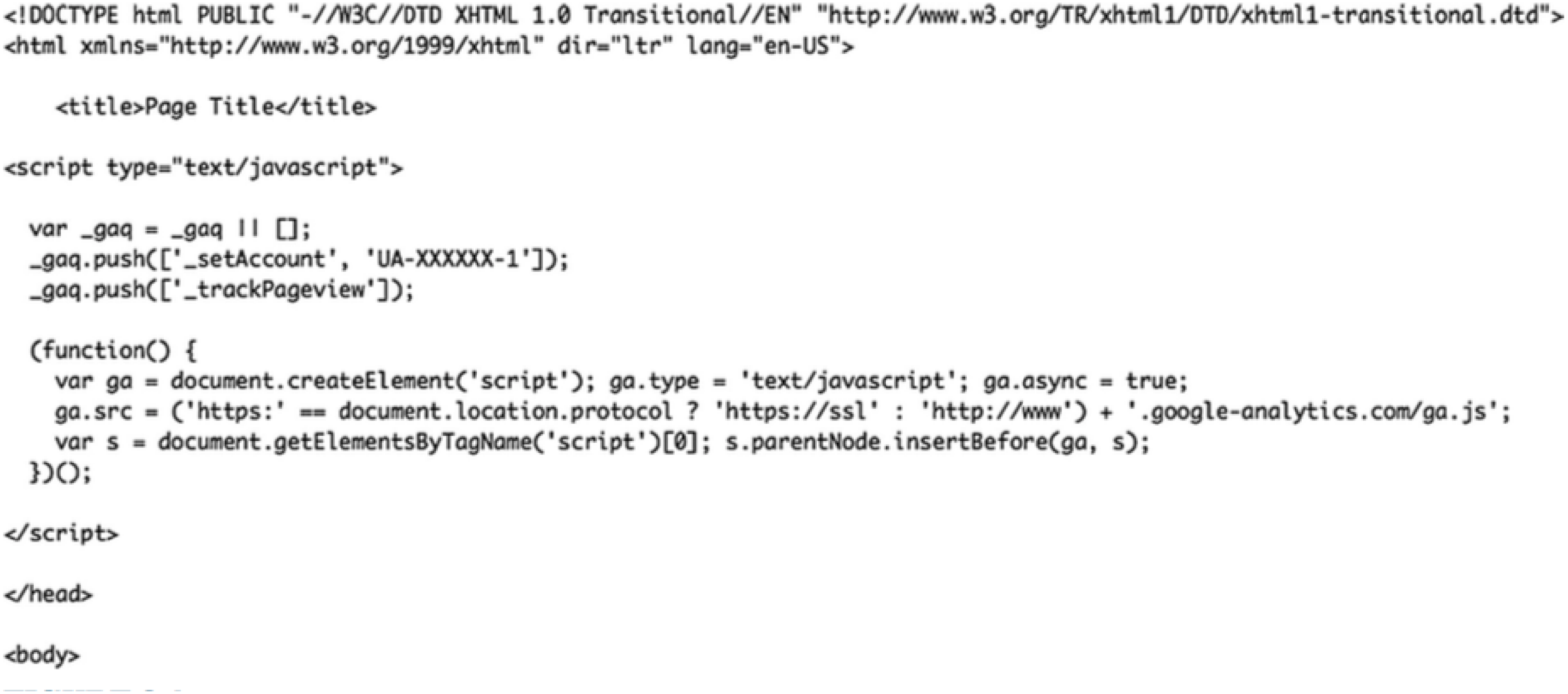
Page tagging

A cookie is also placed on the user's computer

- Allows the analytics tool to determine if its already recorded data from the user
- Allows page to page user activity tracking
- On-page user activity tracking also possible

Data collection

Google Analytics tracking code



```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" dir="ltr" lang="en-US">

    <title>Page Title</title>

<script type="text/javascript">

  var _gaq = _gaq || [];
  _gaq.push(['_setAccount', 'UA-XXXXXX-1']);
  _gaq.push(['_trackPageview']);

  (function() {
    var ga = document.createElement('script'); ga.type = 'text/javascript'; ga.async = true;
    ga.src = ('https:' == document.location.protocol ? 'https://ssl' : 'http://www') + '.google-analytics.com/ga.js';
    var s = document.getElementsByTagName('script')[0]; s.parentNode.insertBefore(ga, s);
  })();

</script>

</head>

<body>
```

Web Beacons

Used to track customer behavior across different websites

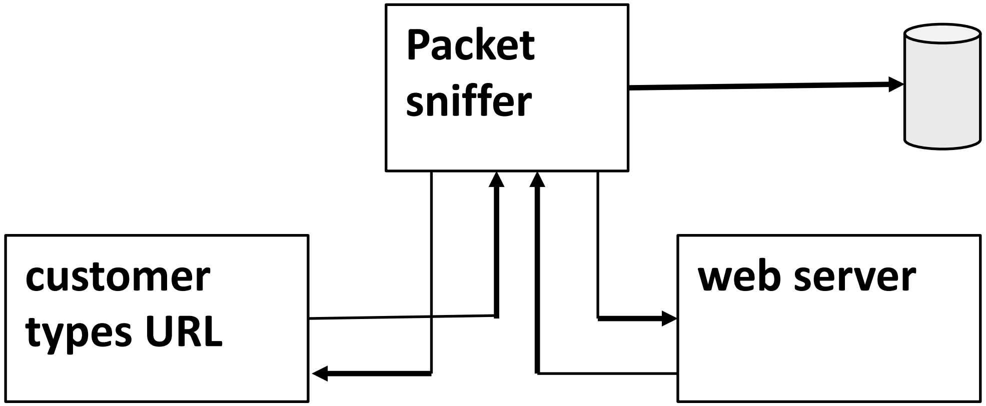
Can be used to measure banner impression and click throughs

Can answer the following question, “How are banner ads performing across multiple websites”

Packet sniffing

Mostly used for multivariate testing

No need to tag pages as all information goes through packet sniffers



Packet sniffer
customer types URL
web server