MMM006-340161 Web Analytics
Programme: Data Engineering

JACOBS
UNIVERSITY

Fall Semester 2018

Lecture 1: Introduction to Web Analytics

Instructor: Dr. Bendick Mahleko

b.mahleko@jacobs-university.de

Objectives

Objectives:

- Disseminate administrative information about the course
- Course outline
- Schedule
- Teaching and Evaluation methods

Motivation

Suppose we own an online retail that sells sport shoes

Questions

- How do potential customers find our store?
- Is it through search engines or through other sites (referrals)?
- How can we know?
- What might it mean if they come to our site and immediately leave?
- Do customer actions call us to action?
- We need to understand why customers behave the way they do

Web Analytics helps to answer some of these questions

Definition

Web Analytics is the collection, measurement, analysis and reporting of websites data for purposes of understanding and optimization.

Four pillars of Web Analytics can easily be identified

- Data collection
- Measurement
- Analysis
- Reporting

Web Analytics deals with Internet customer interaction data (trace data)

Trace data is data left behind indicative of human behavior

Background

Based on an approach that emphasizes the outward behavioral aspects of thought

In Web Analytics behaviorism emphasizes observed behaviors without discounting inner aspects such as context and attitudinal characteristics

Primary proposition

- All things that people do are behaviors
- Focuses primarily on only what the observer can see or manipulate
- Seeks to understand events in terms of behavioral criteria
- Behaviorist research demands to behavioral evidence – important in Web Analytics

Background

Behavioral approach focuses on somebody doing something in a situation

The derived research questions are as follows:

- who (actors)
- what (behaviors)
- when (temporal)
- where (context)
- why (cognitive)

Background

Behaviors can be classified into three general categories

- Behaviors are something that can be detected and therefore recorded
- Behaviors are an action or goal-driven event with some purpose other than the observed action
- Behaviors are reactive responses to environmental stimuli

Web Analytics focuses on descriptive observation and logging the behaviors as they would occur in a user-system interaction episode

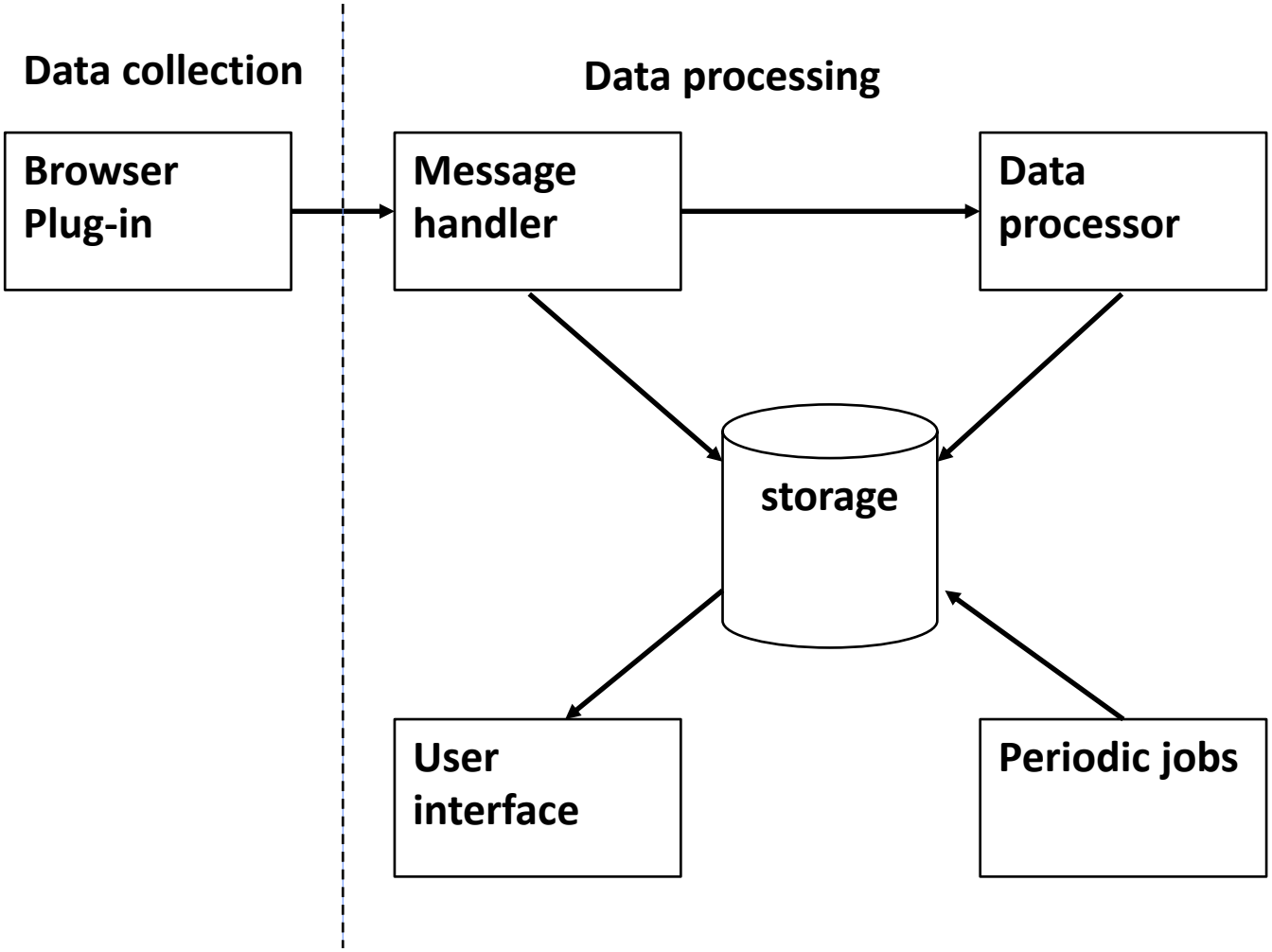
Background

Ethograms are used when studying behavioral patterns in Web Analytics

Ethogram - a taxonomy or index of the behavioral patterns that a user exhibits

State	Description
View results • With scrolling • Without scrolling	User viewed / scrolled one or more pages • User scrolled results page • User did not scroll results page
Browser	User opened, closed or switched browsers
…	…
…	…

Basic overview



Data collection

Cookies and IP addresses

- Cookies are small text files ~4 Kb
- Created on the user's computer
  (a) session cookies are deleted once visitor leaves site
  (b) persistent cookies – remain on visitor's computer

- Cookies help identify new visitors and returning visitors
- IP addresses identify from where a user is accessing a site

Data collection

JACOBS UNIVERSITY

Log files

A log file entry on a server contain the following information

- IP address of user
- Browser type
- Operating system
- Accessed content
- Data and time
- Size of data set

Data collection

JACOBS UNIVERSITY

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

Data collection

JACOBS UNIVERSITY

Log files continued ...

Examples of log analysis tools

- AWStats – (https://sourceforge.net/projects/awstats/)
- Sawmill (https://www.sawmill.net/)
- ...

Data collection

Log files continued ...

Disadvantages of log files

- If the user gets content from cache, the log entry is not recorded
- Log files grow rapidly big and may need to be purged from time to time – makes it difficult to do historical analysis
- Can be complex to implement
- ...

Data collection

JACOBS UNIVERSITY

Page tagging

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

Data collection

Page tagging continued ...

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

Example

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>
```

Pros and cons of page tags

JACOBS UNIVERSITY

Pros	Cons
Breaks through proxy & caching servers, thus providing more accurate session tracking	Must modify webpages to add tags to collect data
Tracks client-side events	Setup errors lead to data loss
Captures client-side ecommerce data. Server side access can be problematic	Firewalls can mangle or restrict tags
Collects & processes data in near real time	Cannot track bandwidth or completed downloads. Tags are set when the page or file is requested
Allows the vendor to perform program updates for you	Cannot track search engine spiders. Robots ignore page tags

Pros and cons of logfile analysis

JACOBS UNIVERSITY

Pros	Cons
Automatic data collection. Does not require page changes	Proxy & caching inaccuracies. If a page is cached, no record is logged on web server
Historical data can be reprocessed	No event tracking
No firewall issues to worry about	Requires own team to perform program updates
Can track bandwidth and completed downloads and can differentiate between completed & partial downloads	Requires own team to perform data storage & archiving
Tracks search engine spiders & robots	Cannot track search engine spiders. Robots ignore page tags
Tracks legacy mobile visitors	Robots inflate visit counts and this can be significant