Web Log & Clickstream

Michel Bruley
WA - Marketing Director

March 2012

Extract from various presentations: Levene, Nagadevara, Waterson, Teradata Aster, …

Definitions

Web Analytics as defined by Web Analytics Association :

" Web Analytics is the measurement, collection, analysis and reporting of Internet data for the purposes of understanding and optimizing Web usage."

Clickstream as defined by Internet Advertising Bureau (IAB) :

"*The electronic path a user takes while navigating from site to site, and from page to page within a site. It is a comprehensive body of data describing the sequence of activity between a user's browser and any other Internet resource, such as a Web site or third party ad server*"

What is a web log?

A record of a visit to a web page

- Visitor (IP address)
- URL
- Time of visit
- Time spent on a page
- Browser used
- Referring URL
- Type of request
- Reply code
- Number of bytes in the reply
- etc…

Web log analysis is a simple kind of Web analytics processes that parses a log file from a web server, and based on the values contained in the log file, derives indicators about who, when, and how a web server is visited.

Usually reports are generated from the log files immediately, but the log files can alternatively be parsed to a database and reports generated on demand.

What is a clickstream?

A record of a path through web pages

- Visitor (IP address)
- URL
- Time of visit
- Time spent on a page
- Browser used
- Referring URL
- Type of request
- Reply code
- Number of bytes in the reply
- Next URL
- etc . . .

A clickstream is the recording of the parts of the screen a computer user clicks on, while web browsing or using another software application.

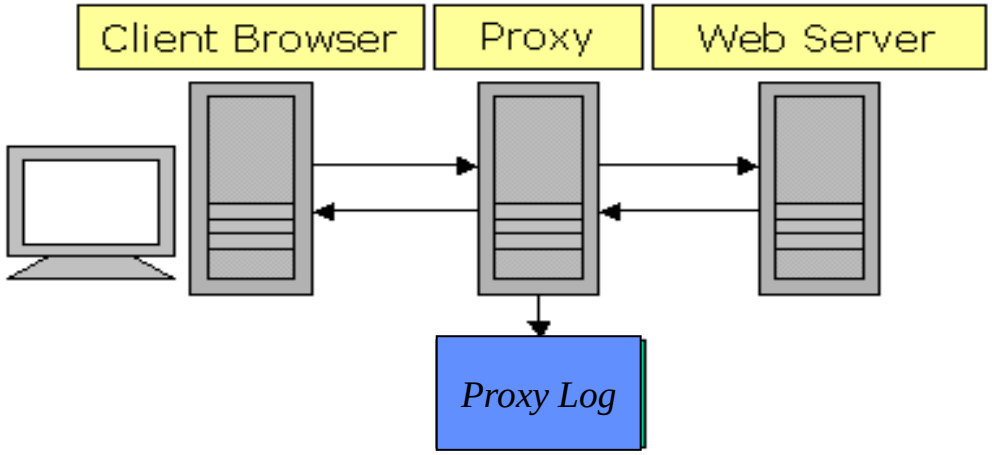
As the user clicks anywhere in the webpage or application, the action is logged on a client or inside the web server, as well as possibly the web browser, router, proxy server or ad server. Clickstream analysis is useful for web activity analysis, software testing, market research, and for analyzing employee productivity.

A small observation on the evolution of clickstream tracking: Initial clickstream or click path data had to be gleaned from server log files.

Web Log: where do they come from?

Servers

- **Done on most web servers**
- **Standard formats**



Clients

- **Browsers, loggers on client machine**
- **Must send data back**

Proxies

- **Similar to servers**
- **Hang out in between client and server**

Why are web logs relevant?

- **Lots of data available**
  - **Quantitative analysis is much more fun!**
- **User behavior, patterns**
  - **Real users, tasks**
  - **Or at least more realistic users and tasks**
- **Leaving the usability lab**
  - **Testing effect**
- **Fast, easy, cheap**
  - **Automatic or almost-automatic**

Some questions on Web usages

- How has information been accessed?
- How frequently?
- What's popular? What's not?
- How do people enter the site? Exit?
- Where do people spend time?
- How long do they spend there?
- How do people travel within the site?
- Who are the people visiting?

Some others questions

- **Structural**
  - **What information has been added, deleted, modified, moved?**
- **Usage + Structural**
  - **What happens when the site changes?**
  - **Does navigation change?**
  - **Does popularity change?**
  - **What about missing data?**

How do you analyze web logs?

- **Data Mining: task or intent unknown**
  - **"Automated extraction of hidden predictive information from (large) databases"**
  - **Server log analysis**

What are people doing?

- **Remote Usability Testing: task or intent known**
  - **Similar to traditional lab usability testing**
  - **Clickstream analysis**

How well does the site support what people are doing?

Web Log File Basic Analysis

- **Gives statistics such as**
  - **number of hits**
  - **average hits per time period**
  - **what are the popular pages in your site**
  - **who is visiting your site**
  - **what keywords are users searching for to get to you**
  - **what is being downloaded**
- **Log data does not disclose the visitor's identity**

Identification of User

- By IP address
  - Not so reliable as IP can be dynamic
  - Different users may use same IP
- Through cookies
  - Reliable but user may remove cookies
  - Security and privacy issues
- Through login
  - Users have to register

Sessionising

- **Time oriented (robust)**
  - **By total duration of session**
    - **not more than 30 minutes**
  - **By page stay times (good for short sessions)**
    - **not more than 10 minutes per page**
- **Navigation oriented (good for short sessions and when timestamps unreliable)**
  - **Referrer is previous page in session, or**
  - **Referrer is undefined but request within 10 secs, or**
  - **Link from previous to current page in web site**

Mining Navigation Patterns

- **Each session induces a user trail through the site**
- **A trail is a sequence of web pages followed by a user during a session, ordered by time of access**
- **A pattern in this context is a frequent trail**
- ***Co-occurrence* of web pages is important, e.g. *shopping-basket* and *checkout***
- **Use a Markov chain model**

Starting with Raw Click Stream Data

96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-log.enquisite.com/d.js?id=a1a3af-ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"

Finding Referring Channels

96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-log.enquisite.com/d.js?id=a1a3af-ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"

Identifies the Customer

Identifies the customer by correlating Cookies captured during customer interactions …

referrer=http://www.facebook.com/

sc=ar78to0wdxq143lr9p9weh6yxy8318sa
pc=pgys63w0xgn102in8ms37wka8quxe74e
ac=bd76aad17448wgng0679a044cfda00e005b130000

Persistent and Network Cookies remain unchanged

Creates a Session Definition

96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-log.enquisite.com/d.js?id=a1a3af-ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"

Tracking of Customer Interaction Behavior During the Session

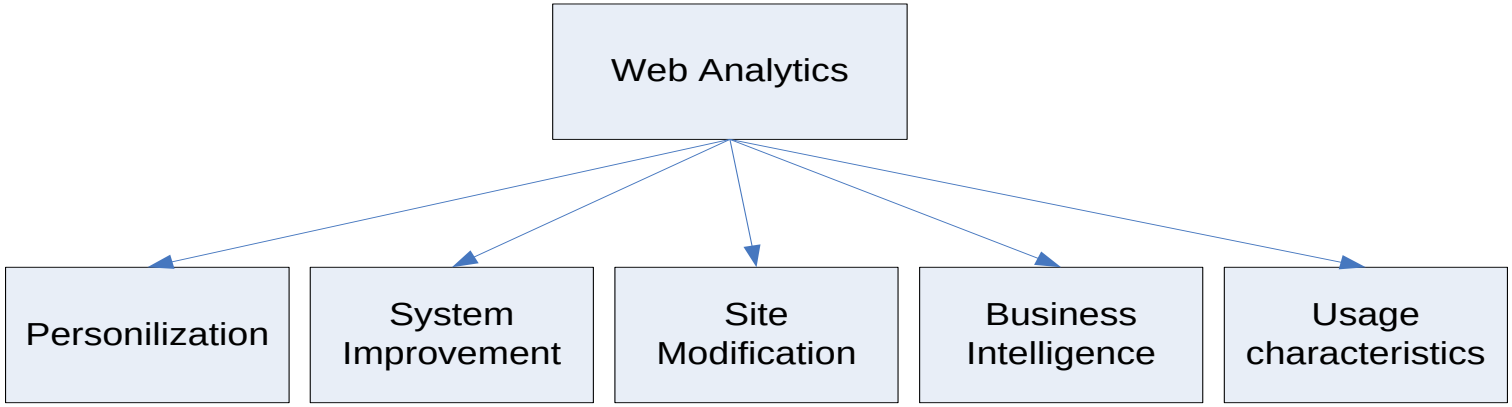
Some Classic Measures

- **Clicks: The interaction between the user and the web server is measured by the click of a mouse**
- **Visits: The number of times a user visits a specific web site. Every new session is counted as a new visit**
- **Hits: Total number of server requests serviced by the server**
- **Exits: Site exits, counted by site inactivity for more than 30 minutes**
- **Unique Visitors: A Unique User who accesses the site in a specified period of time.**
- **Repeated Visitor: The average number of times a user returns to a site over a specific time period**
- **Page views: The view of any page by the user. A page may contain text, images, and other online elements and may be statically or dynamically generated and could contain single or multiple frames or screens**
- **Sessions: IAB defines it to be an "A sequence of Internet activity made by one user at one site. If a user makes no request from a site during a 30 minute period of time, the next content or ad request would then constitute the beginning of a new visit "**
- **Unique authenticated visitors: A unique visitor who logs on to a site via a registration method using his/her user id and password**

Some Classic Metrics

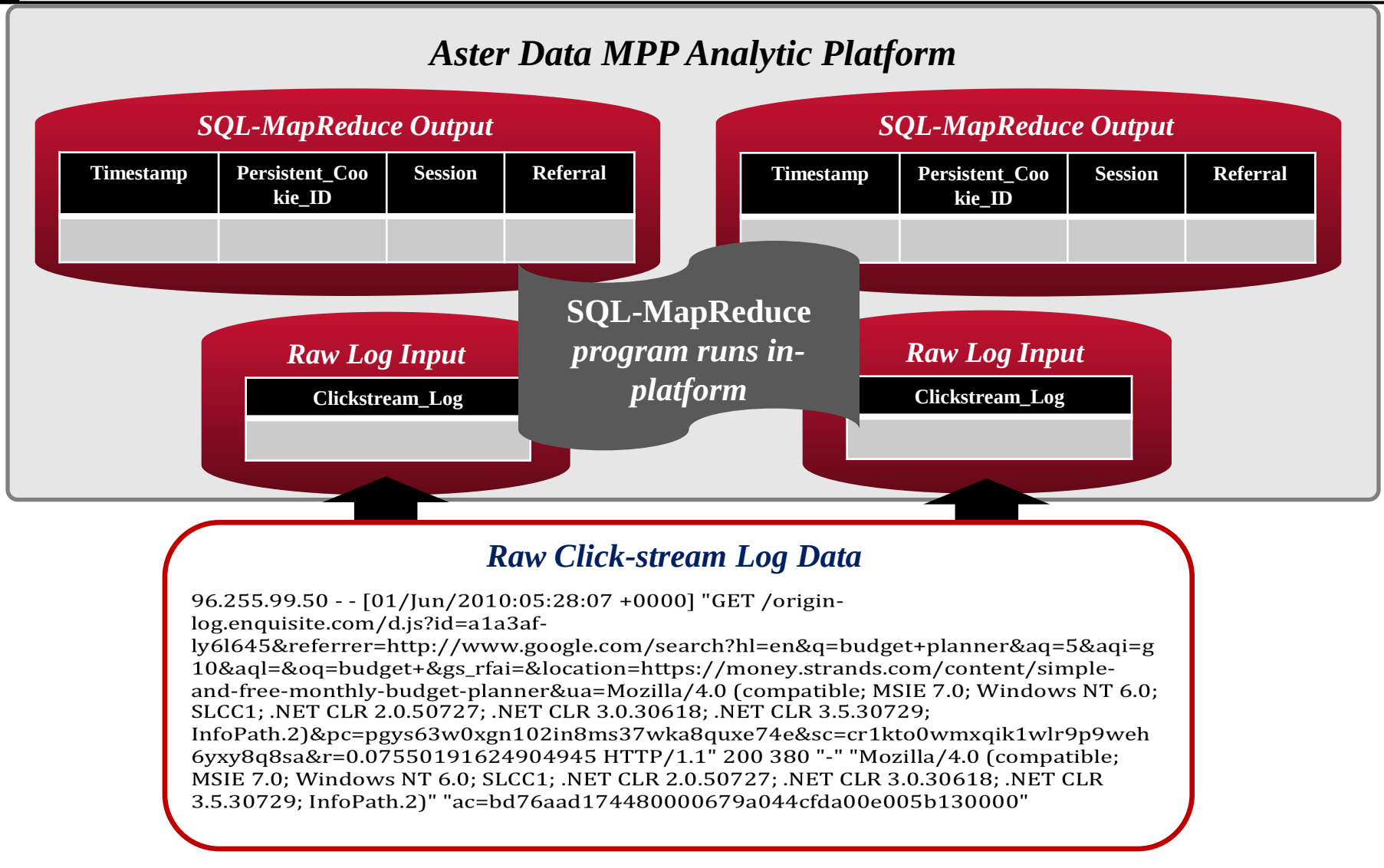
- **Page views per visit: Average number of page views per visit**
- **Page views per session: Average number of page views per session**
- **Page views per hour/day: Average number of page views per hour/day**
- **Clicks per session: Average number or clicks per session**
- **Clicks per hour: Average number of clicks per hour**
- **Time between clicks: The average duration of time spent between two clicks**
- **Hits per hour: Average number of hits to the web server per hour**
- **Busy hour of the day: The highest number of hits to the web server in a particular hour of a day**

Information from Web Analytics



How many visitors visit the page daily?
Who are the regular visitors?
What percentage of the visitors to the page are registered users?
What are the top pages that are visited on the web page?
What is the average visit time on the website?
How often does the visitor return to the site?
What is the average page depth of a visitor?
What is the geographic distribution of users of the website?

Teradata Aster Interpret Web Log Data



Raw Click-stream Log Data

96.255.99.50 - - [01/Jun/2010:05:28:07 +0000] "GET /origin-log.enquisite.com/d.js?id=a1a3af-ly6l645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g10&aql=&oq=budget+&gs_rfai=&location=https://money.strands.com/content/simple-and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)&pc=pgys63w0xgn102in8ms37wka8quxe74e&sc=cr1kto0wmxqik1wlr9p9weh6yxy8q8sa&r=0.07550191624904945 HTTP/1.1" 200 380 "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"

Teradata Aster Competitive Advantage

Internet Use Cases	Financial Services & Insurance Use Cases	Retail Use Cases	Media & Information Services Use Cases
• Social networking graph analysis • Crowd-sourcing • Virality analysis • Content targeting • Advanced click-stream analysis	• Real-time fraud & link analysis • Tick data analysis • Trading surveillance • Multi-variate pricing analysis for insurance • Behavior pattern matching	• Digital marketing attribution • Online consumer behavior/patterns • Advanced click-stream analysis • Online targeting for personalization/ recommendations	• Predictive and granular forecasting • Advanced click-stream analysis • Digital media consumer micro-targeting • Ad optimization

"Analytics themselves don't constitute a strategy, but using them to optimize a distinctive business capability certainly constitutes a strategy."

- T. Davenport & J. Harris, Competing on Analytics: The New Science of Winning