Web Analytics

Web Traffic Data Sources & Vendor Comparison

A whitepaper by Brian Clifton in conjunction with Omega Digital Media Ltd

Updated May 2008

omega digital media

ANALYTICS AUTHORIZED CONSULTANT Google

Web Traffic Data Sources & Vendor Comparison

Table of Contents

Preface

When it comes to benchmarking the performance of your web site, web analytics is critical. The industry that started in 1995 for webmasters, is now rapidly evolving so that it is almost a mainstream part of digital marketing. This whitepaper compares the different data collection techniques available, shows the competitive landscape for web analytics vendors and illustrates the major milestones of the industry over the past years.

About the Author

Brian Clifton (PhD), is an internationally recognized search marketing and web analytics expert who has worked in these fields since 1997. A respected speaker at conferences, including Search Engine Strategies, Internet World, eMetrics and ad:tech, Brian is also the author of a number of industry whitepapers and recently published the book entitled Advanced Web Metrics with Google Analytics.

In 2005 Brian joined Google as Head of Web Analytics for Europe, Middle East and Africa. Defining the strategy for adoption and building a pan-European team of product specialist for operational support, the Google Analytics product became the market leader for the world's largest online advertisers within two years.

Brian is now Senior Strategist for Omega Digital Media - a company specialising in search integration and conversion marketing for European clients.

If you have comments about this document, add your views at: www.advanced-web-metrics.com/blog/recommended-reading.

Web Traffic Data Sources & Vendor Comparison

Once you have decided that you need to analyse your web site visitor traffic, the next most important step, before evaluating a vendor, is to determine exactly which data it is you are going to analyse.

Different Visitor Data Collection Methods

By far the most common form (estimated as 99%+ of all accounts) of collecting web visitor data are Page Tags and Logfiles.

Page Tags refer to data collected by a visitors' web browser, achieved by placing “beacon” code on each page of your site. Often it is simply a single snippet (tag) of code referencing a separate javascript file – hence the name. Some vendors also add multiple custom tags to set/collect further data. This type of technique is known as **client-side data collection**.

Logfiles refer to data collected by your web server, which is independent of a visitors' browser. By default, all requests to a web server (pages, images, pdf's etc) are logged to a file – usually in plain text. This type of technique is known as **server-side data collection**.

Logfile analysis was historically the way to analyse web site visitor behaviour. Web server logfiles are readily available, hence site owners simply purchased the software to analyse their logfiles. However page tagging has become very popular in recent years.

It is important to note that both techniques, when considered in isolation, have their limitations. Table 1 summarises the differences and shows that by combining both, the advantages of one counters the disadvantages of the other. This is known as a HYBRID method. That is, combining both web logs with page tags.

The main reason that page tag techniques are now flourishing, is that they allow analysis to be outsourced, commonly referred to as a "Hosted" solution. That is, the data is collected and processed away from your organisation, saving you (the web site owner) the IT worries of configuring and maintaining your own software as well as the storing and archiving of collected data.

Whilst a Hosted solution may be your best option for business reasons, bear in mind most hosted solutions are based on page tags only. A common myth is that page tags are technically superior to other methods, but as Table 1 shows, that depends on what you are looking at. Only a hybrid solution can provide a complete analysis of your web site visitor behaviour. Because of their complexities, most hybrid solutions are software based. However a small number of vendors can offer a hosted hybrid solution.

Other data collection methods

Note that although logfile analysis and page tagging are the most prolific ways to collect web visitor data, they are not the only methods. Network Data Collection devices or "packet sniffers" gather web traffic data from routers into 'black box' appliances. Possibly because of implementation complexities/cost, only a couple of vendors are known to use the NDC method.

Another technique is to use a web server API/Loadable Module (also known as a plugin, though not strictly correct). These are programs that extend the capabilities of the web server. For example, enhancing and/or extending the fields that are logged.

Costs of Data Collection

The price of hard disk space and bandwidth is now so cheap that some page tag vendors will collect data for you for free. These include Google Analytics, Microsoft adCenter Analytics and Yahoo IndexTools.

Of course there is a resource cost for you to consider in terms of implementation of these free tools – even if you chose a DIY route. Other paid-for page tag vendors charge an implementation fee plus data collection fees by volume i.e. X pageviews per month.

Using server-side web analytics tools to analyse logfiles liberates you from pageview fees. However, the true cost of ownership of running and managing your own licensed software also needs to be considered. For example, is a dedicated server required? Software upgrades, logfile maintenance, archiving etc. all need to be managed by your IT team and this cost should be included.

Table 1 – Methodology Pros and Cons

Page Tagging v Logfile Analysis

Page Tagging

Advantages

- Breaks through proxy and caching servers
  - provides more accurate session tracking
- Track client side events
  - JavaScript, Flash, web 2.0
- Client-side capture of e-commerce data
  - server-side access can be problematic
- Visitor data can be collected/processed in near real-time
- Program updates performed for you by the vendor
- Data storage and archiving performed for you by the vendor

Disadvantages

- Setup errors lead to data loss
  - If you make a mistake with your tags, data is lost and you can not go back and re-analyse
- Firewalls
  - can mangle or restrict tags
- Cannot track bandwidth or completed downloads
  - Tags are set when the page/file is requested not when the download is complete
- Cannot track search engine spiders
  - robots ignore page tags

Logfile Analysis

Advantages

- Historical data can be reprocessed easily
- No Firewall issues to worry about
- Can track bandwidth and completed downloads
  - also differentiate completed and partial downloads
- Track search engine spiders/robots by default
- Track mobile visitors by default

Disadvantages

- Proxy/caching inaccuracies
  - If a page is cached, no record is logged on your web server
- No event tracking (javascript, Flash, web v2.0)
  - no Javascript, flash, web v2.0 tracking
- Program updates performed by your own team
- Data storage and archiving performed by your own team

Cookie Considerations

Cookies are small text messages that a web server transmits to a web browser so that it can keep track of the user's activity on a specific web site. The visitors' browser stores the cookie information on the hard drive so when the browser is closed and reopened at a later date, the cookie information is still available. These are known as *persistent cookies*. Cookies that only last a visitors' session are known as *session cookies*.

The main purpose of cookies is to anonymously identify users for later use – most often a visitor ID number. This can be used for example to determine how many first time or repeat visitors a site has received, how many times a visitor returns each period and what is the length of time between visits.

Web servers can also use cookie information to present custom web pages i.e. a returning visitor may be shown different content than a first time visitor. If you register or login to a service, other cookie information may be used to personalise the information e.g. Welcome back Brian.

There are two types of cookies: **first-party** and **third-party**. A first-party cookie is one created by the web site you are currently visiting. A third-party cookie is sent from a web site different from the one you are currently viewing. The idea is that the transfer of cookie information takes place behind the scenes without the user having to know/worry about it. However this does mean cookies have implications relevant to a user's privacy and anonymity on the Internet.

From a web analytics point of view, cookie information is very important. The general best practice consensus is that vendors should only set and process first-party cookies. The rationale is that many anti-spy programs and firewalls exist that will block third party cookies by default, therefore mangling the collected analytic data. The interpretation is that third-party cookies make behavioural information available to third parties, that the web visitor is either not aware of or not consented to i.e. infringing on privacy.

End-users are also becoming much more 'cookie savvy' and will often delete cookies manually or set their browser settings so as to reject third party cookies automatically. Recent studies have indicated that as many as 30% of users delete cookies within 30 days.

Cookie facts:

- Cookies are small text files, stored locally, that are associated with visited web site domains.
- Cookie information can be viewed by users of your computer, using Notepad or a text editor application.
- There are two types of cookies – first-party and third-party: A first-party cookie is one created by the web site domain that a visitor requests directly either by typing in the URL into their browser or following a link. A third-party cookie is one that operates in the background and is usually associated with advertisements or embedded content that is delivered by a third party domain not directly requested by the visitor.
- For first-party cookies, only the web site domain setting the cookie information can retrieve this data. This is a security feature built into all web browsers.
- For third-party cookies, the web site domain setting cookie can also list other domains allowed to view this information. The user is not involved in the transfer of third-party cookie information.
- Cookies are not malicious and can't harm your computer. They can be deleted by the user at any time.
- Cookies are no larger than 4 kilobytes.
- A maximum of 50 cookies are allowed per domain for the latest versions of IE7 and Firefox 2. Other browsers may vary (Opera 9 currently has a limit of 30).

Table 2 – Competitive Landscape

Note: this is a working document. If you are vendor (or know of one), that isn't on the list, simply send the details for inclusion.

Notes:*

- *Data Collection Methods:*

  ***SS** –uses server-side collected data e.g. web server logfiles, though may also be web server API*

  ***CS** –uses client-side collected data e.g. page tags usually written in javascript in conjunction with a pixel gif. This can be a 'tags into logs' approach or an interaction between active collection servers and page tags to control and organise data collection on the fly (i.e. dynamic tags). May also be web server API*

  ***Hybrid** – combines server-side and client-side collected data to effectively augment/fortify data therefore reducing the inaccuracies of only using either/or method. For Hybrids, some vendors use page tagging to collect client-side data into cookies, which are then logged into the web server log files i.e. "cookie-fortified logs". Other vendors use a web server plugin API to effectively do the same thing, but replace the logging capabilities of the web server (allows logs to be collected externally). Hence both techniques are simply labelled as Hybrid.*

- *S/ware (S) and/or Hosted (H): Can the client buy the software license and setup/run as they wish, or is it a hosted solution controlled by the vendor on a lease agreement (usually charged by volume i.e. page views per month). Network Data Collection (NDC) devices or "packet sniffers" are also listed here.*

- *Confirmed by: This is simply a knowledgeable person (vendor, client, forum user) that has confirmed the Data Collection Method.*

- *Comments: Comments added by Brian Clifton to augment data. Comments are <u>not</u> a feature list or sales pitch and are purely for information purposes. If you wish to add/change information, please email the author with the following considerations:*

  - *Comments are limited to 300 characters*
  - *No superlatives, no sales pitch, no pricing info*
  - *The author has the right to reject or amend comments*
  - *I am particularly interested to hear from UK/EU vendors that have achieved technology firsts*

Thanks to all that posted responses at tech.groups.yahoo.com/group/webanalytics and from personal contacts.

Web Traffic Data Sources & Vendor Comparison

Vendor Name	Origin	DOB	Data Collection			S/ware (S) and/or Hosted (H)	Confirmed by	Comments
			SS	CS	Hybrid			
Clickstream.com	UK	1999	✓	✓	✓	N/A	Rufus Evison	*****First Hybrid 1998***** API. Allows logs to be collected externally. Similar in principal to Visual Sciences. Solely a data collection/technology provider i.e. not a reporting package. Hybrid method developed by Green Cathedral Plc which Clickstream demerged from (1999).
Clicktracks.com *Now part of J.L.Halsey*	US	2002	✓	✓	-	S,H	John Marshall	Windows only. Requires desktop application in addition, uses 3rd party cookies
Coremetrics.com	US	1999	✓	✓	✓	H	Frank Lombos	Uses 1st party cookies.
DeepMetrix.com *Now Microsoft adCenter Analytics*	CA	1996	✓	✓	✓	S,H		Hosted solution uses page tags, software (Windows only) uses page tags + server logs. Ships with MSDE, though MS SQL required for large installations. Hosted is page tags only.
evisitanalyst.com	UK	March 2002	-	✓	-	S,H	Adam Hulme	Uses 3rd party cookies. Able to track 'back button' activity. Hosted is page tags only
Fireclick.com	US	1999	-	✓	-	H	Xavier Casanova	Page tags only
Google Analytics *Formerly Urchin*	US	2005 *1997*	✓	✓	✓	S,H	Jason Senn	Multi-platform, hybrid since Jun 2002 Hosted is page tags only. Only 1st party cookies. Software uses augmented logfiles i.e. page tags + server logs to produce 'cookie-fortified' logs.
HitMatic.com	UK	1999	-	✓	-	H		Page tags only
IBM SurfAid *Now Coremetrics*	US	1998	✓	✓	✓	H	Michael Horn Michael Nichols	Uses 1st or 3rd party cookies.
IndexTools *Now part of Yahoo*	HU	Jun 2000	-	✓	-	H	Dennis R. Mortensen	Page tags only. Uses 1st and 3rd party cookies
InSite	UK	2002	-	✓	-	S,H	Brandt Dainow	Page tags only. Can also track search engine positions.
Instadia.net *Now part of Omniture*	DK	2000	✓	✓	-	H	Anders F. Jorgensen	Hosted solution can also report on Intranet users by piping internal logs directly into Instadia.
Intellitracker.com	UK	1997	✓	✓	✓	H	Satin Dattani	Introduced hybrid 2004.
Moniforce.com	NL	May 2001	✓	✓	✓	S,H, NDC	Katja Graaf	Hosted (page tags only) or hybrid solution supplied as a black box (NDC) appliance. Hybrid since Q3 2004. Uses 1st party cookies.
mtracking.com	UK	2002	-	✓	-	H		Page tags only
Nedstat.com	NL	1996	-	✓	-	S,H		Page tags only. Uses 3rd party cookies.

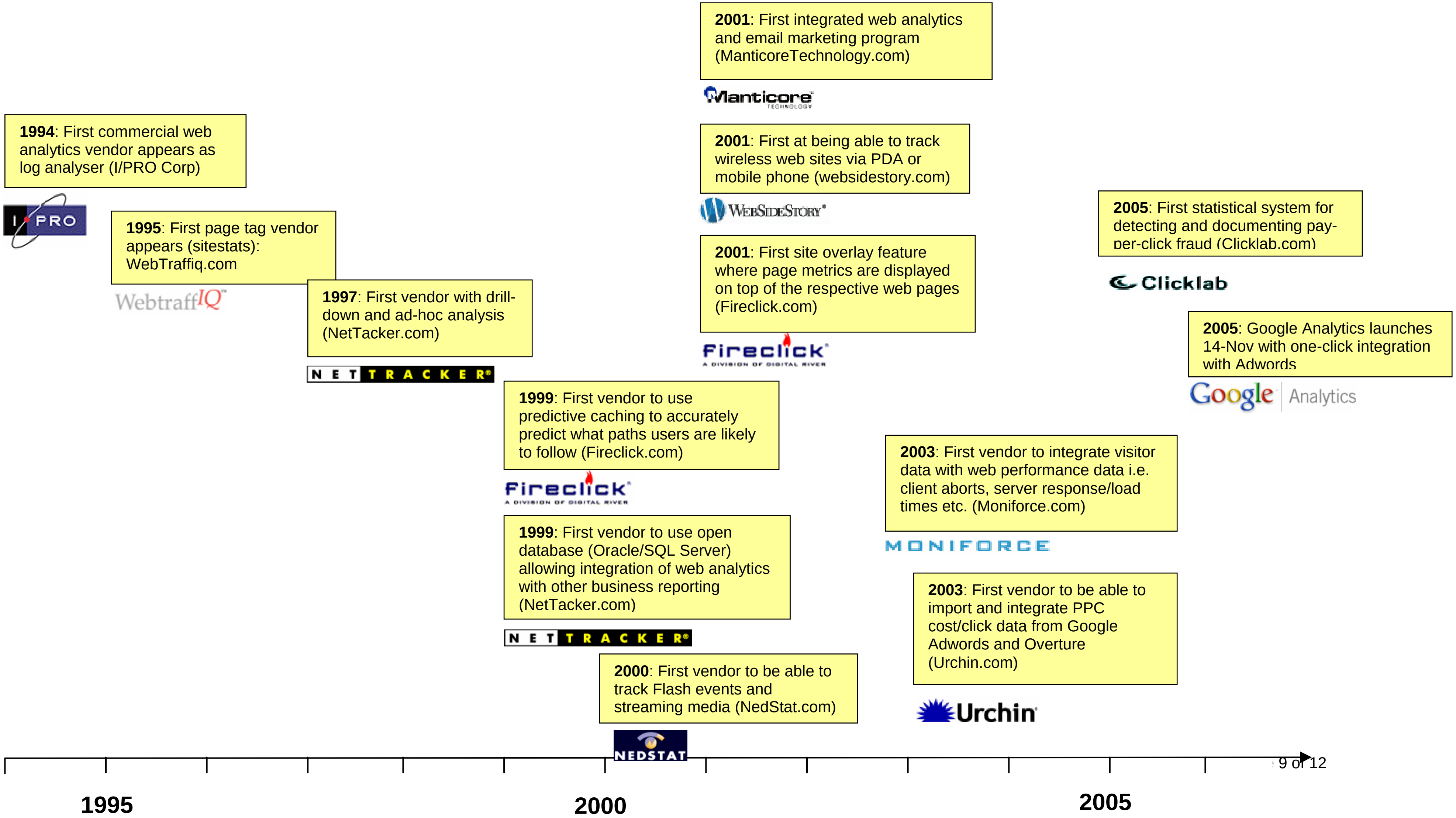
Web Traffic Data Sources & Vendor Comparison

Vendor Name	Origin	DOB	Data Collection			S/ware (S) and/or Hosted (H)	Confirmed by	Comments
			SS	CS	Hybrid			
NetTracker.com *Now part of Unica*	US	1996	✓	✓	✓	S,H	Akin Arikan	Multi-platform, hybrid since Oct 2004. Uses augmented logfiles i.e. page tags + server logs to produce 'cookie-fortified' logs. Can provide hosted hybrid solution. Uses 1st or 3rd party cookies.
Omniture.com	US	2002	-	✓	-	H	Matt Belkin	Page tags only. Uses 1st or 3rd party cookies.
Redeye.com	UK	1997	-	✓	-	H	Bertie Stevenson	Page tags only. Main technique is identifying visitors by a login where possible.
Site Census *Formerly RedSheriff*	AU	1996	?	?	?			
SageMetrics.com *Now part of Blue Freeway*	US	1997	✓	✓	✓	H	Benoit Droulez	Hybrid from 2001. Possibility to merge external data sources (registration, sales, etc.) with web traffic. Can use 1st or 3rd party cookies
Sawmill.co.uk	US	1997	✓	-	-	S	Les Ferrington	Logfile analysis only. Multi-platform, multi-logfile - not just web analytics
Site-intelligence.co.uk	UK	2000	✓	✓	-		David Pool Guy Evans	Uses 1st party cookies
speed-trap.com	UK	Dec 1999	-	✓	-	S,H	Malcolm Duckett	Uses 'active' page tags (javascript or java) i.e. collection server conducts a dialog with the page tags which sends the data back. Has OEM (white label) solutions. Can integrate with other JDBC sources
TeaLeaf	US	1999	✓	✓	✓	NDC		Sniffs all input at the TCP/IP level
VisualSciences.com *Now part of Omniture*	US	Sep 2001	✓	✓	✓	S,H	Jim MacIntyre	Hybrid from Oct 2001. Supports page tags and/or web server API as well as log files and/or ODBC sources. Can provide hosted hybrid solution.
WebAbacus.co.uk *Now part of Foviance*	UK	?	✓	✓	✓	S,H	Ian Thomas	
WebTrends.com	US	1995	✓	✓	✓	S,H	Barry Parshall	Software (Windows only) processes server logs + page tags. Hybrid introduced Apr 2004 (v7.0). Software licensed by page views. Can provide hosted hybrid solution (Jan 2005). Uses 1st or 3rd party cookies.
WebSideStory (HBX) *Now part of Omniture*	US	1996	-	✓	-	H	Jay Calavas	Page tags only. Uses 1st party cookies.
Webtraffiq.com *Now part of Moore-Wilson*	UK	1995	✓	✓	✓	(S),H	Marcos Richardson	Software/hybrid can be provided as bespoke solution. Use ROLAP for multi dimensional analysis. Also integrates with ODBC data sources. Hosted is page tags only. Uses 1st party cookies
Xiti.com	FR	2000	-	✓	-	H	Benoit Arson	Page tags only. Uses 1st party and 3rd party cookies

Web Traffic Data Sources & Vendor Comparison

Vendor Timeline of Technology Firsts

Throughout the past decade, vendors have battled it out to develop additional features. This ‘feature war’ was the main differentiator for vendors. However the industry has matured enough to provide a great deal of feature parity between vendors. Major features such as geo-location lookup, cross data segmentation, multi-line trending, Search Engine Marketing are now standard. The below chart highlights some of the key vendors that contributed to the development of these features.



1994: First commercial web analytics vendor appears as log analyser (I/PRO Corp)

1995: First page tag vendor appears (sitestats): WebTraffiq.com

1997: First vendor with drill-down and ad-hoc analysis (NetTacker.com)

1999: First vendor to use predictive caching to accurately predict what paths users are likely to follow (Fireclick.com)

1999: First vendor to use open database (Oracle/SQL Server) allowing integration of web analytics with other business reporting (NetTacker.com)

2000: First vendor to be able to track Flash events and streaming media (NedStat.com)

2001: First integrated web analytics and email marketing program (ManticoreTechnology.com)

2001: First at being able to track wireless web sites via PDA or mobile phone (websidestory.com)

2001: First site overlay feature where page metrics are displayed on top of the respective web pages (Fireclick.com)

2003: First vendor to integrate visitor data with web performance data i.e. client aborts, server response/load times etc. (Moniforce.com)

2003: First vendor to be able to import and integrate PPC cost/click data from Google Adwords and Overture (Urchin.com)

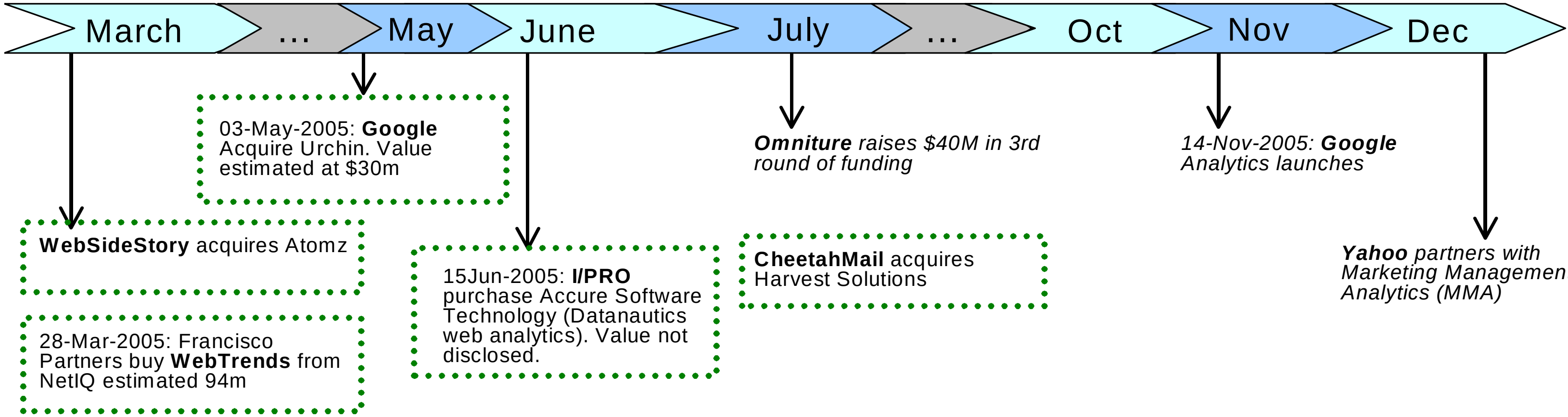
2005: First statistical system for detecting and documenting pay-per-click fraud (Clicklab.com)

2005: Google Analytics launches 14-Nov with one-click integration with Adwords

Web Traffic Data Sources & Vendor Comparison

Vendor Newswires & Significant Events

2005



WebSideStory acquires Atomz

28-Mar-2005: Francisco Partners buy **WebTrends** from NetIQ estimated 94m

03-May-2005: **Google** Acquire Urchin. Value estimated at $30m

15Jun-2005: **I/PRO** purchase Accure Software Technology (Datanautics web analytics). Value not disclosed.

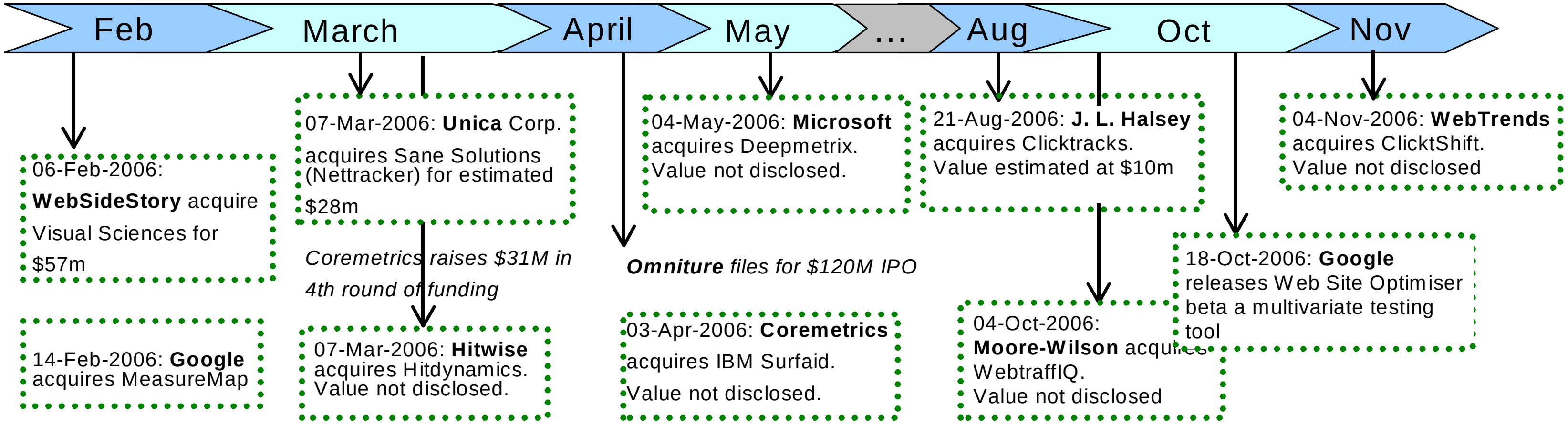
***Omniture** raises $40M in 3rd round of funding*

CheetahMail acquires Harvest Solutions

*14-Nov-2005: **Google** Analytics launches*

***Yahoo** partners with Marketing Management Analytics (MMA)*

2006



06-Feb-2006: **WebSideStory** acquire Visual Sciences for $57m

14-Feb-2006: **Google** acquires MeasureMap

07-Mar-2006: **Unica** Corp. acquires Sane Solutions (Nettracker) for estimated $28m

Coremetrics raises $31M in 4th round of funding

07-Mar-2006: **Hitwise** acquires Hitdynamics. Value not disclosed.

***Omniture** files for $120M IPO*

03-Apr-2006: **Coremetrics** acquires IBM Surfaid. Value not disclosed.

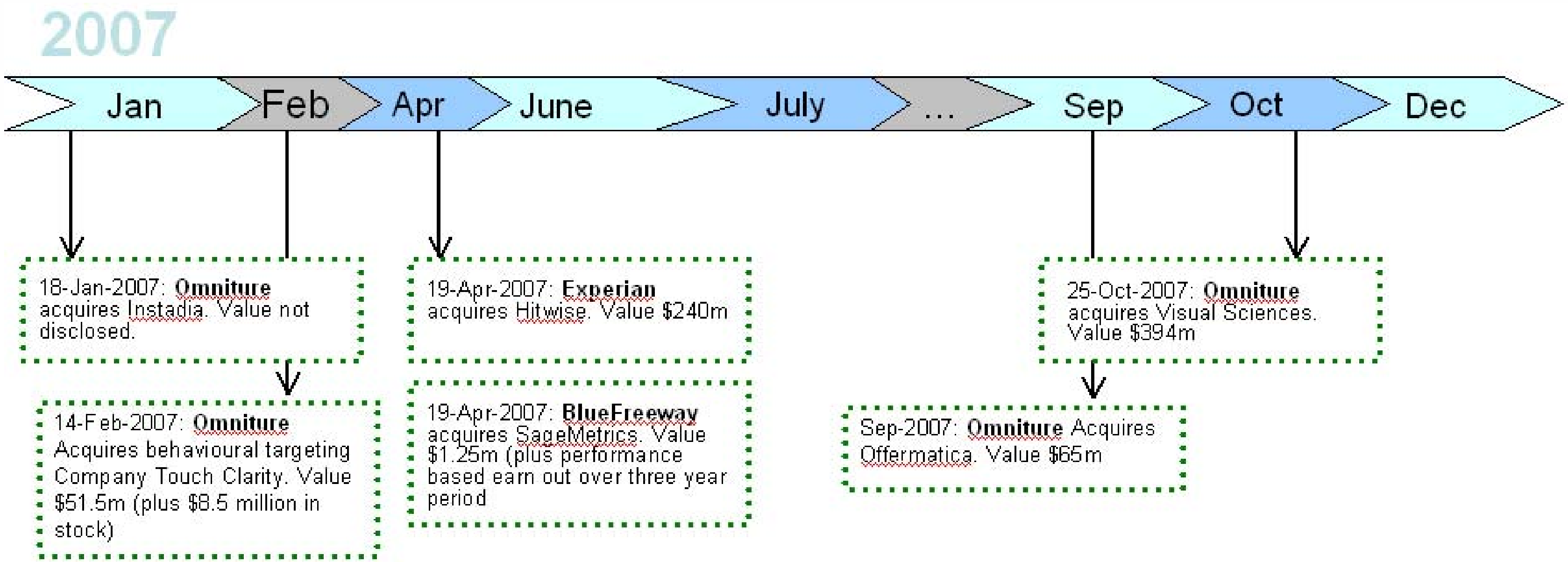
04-May-2006: **Microsoft** acquires Deepmetrix. Value not disclosed.

21-Aug-2006: **J. L. Halsey** acquires Clicktracks. Value estimated at $10m

04-Oct-2006: **Moore-Wilson** acquires WebtraffIQ. Value not disclosed

18-Oct-2006: **Google** releases Web Site Optimiser beta a multivariate testing tool

04-Nov-2006: **WebTrends** acquires ClicktShift. Value not disclosed



18-Jan-2007: **Omniture** acquires Instadia. Value not disclosed.

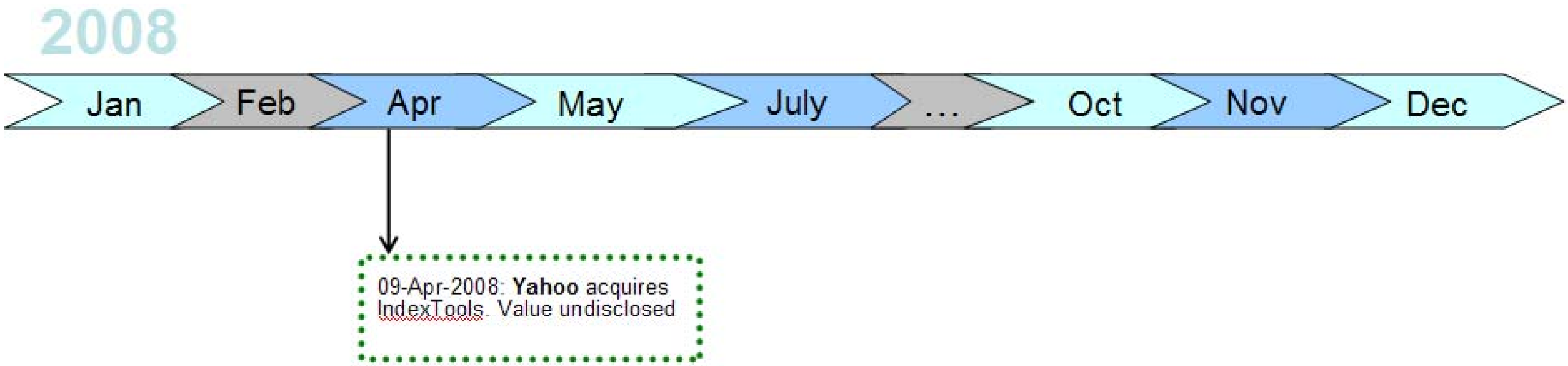
14-Feb-2007: **Omniture** Acquires behavioural targeting Company Touch Clarity. Value $51.5m (plus $8.5 million in stock)

19-Apr-2007: **Experian** acquires Hitwise. Value $240m

19-Apr-2007: **BlueFreeway** acquires SageMetrics. Value $1.25m (plus performance based earn out over three year period

25-Oct-2007: **Omniture** acquires Visual Sciences. Value $394m

Sep-2007: **Omniture** Acquires Offermatica. Value $65m



09-Apr-2008: **Yahoo** acquires IndexTools. Value undisclosed

Further Recommended Reading

Other white papers in this series from Brian Clifton:

Increasing Accuracy for Online Business Growth

> This 14 page document describes the accuracy limitations of on-site web analytics tools and how can you mitigate these and get comfortable with your data. Importantly, it is vendor agnostic. That is, with a best practice implementation of your web analytics tool, you can get very precise visitor data.

How Search Engine Optimisation (SEO) works

> Updated for its 7th year in circulation with over 10,000 downloads, this 16 page document is an excellent primer for anyone wishing to understand the intricacies of SEO.

Web Analytics Data Sources – **this one!**

A list of recommended reading of books and whitepapers is available from advanced-web-metrics.com/blog/recommended-reading