

## Web Analytics

### Web Analytics

Web Analytics is the study of the behavior of website visitors. In a commercial context, web analytics especially refers to the use of data collected from a web site to determine which aspects of the website work towards the business objectives; for example, which landing pages encourage people to make a purchase.

Data collected almost always includes web traffic reports. It may also include e-mail response rates, direct mail campaign data, sales and lead information, user performance data such as click heat mapping, or other custom metrics as needed. This data is typically compared against key performance indicators for performance, and used to improve a web site or marketing campaign's audience response.

Many different vendors provide web analytics software and services.

### Web analytics technologies

There are two main technological approaches to collecting web analytics data. The first method, logfile analysis, reads the logfiles in which the web server records all its transactions. The second method, page tagging, uses JavaScript on each page to notify a third-party server when a page is rendered by a web browser.

### Web server logfile analysis

Web servers have always recorded all their transactions in a logfile. It was soon realised that these logfiles could be read by a program to provide data on the popularity of the website. Thus arose web log analysis software.

In the early 1990s, web site statistics consisted primarily of counting the number of client requests (or hits) made to the web server. This was a reasonable method initially, since each web site often consisted of a single HTML file. However, with the introduction of images in HTML, and web sites that spanned multiple HTML files, this count became less useful. The first true commercial Log Analyzer was released by IPRO in 1994.

Two units of measure were introduced in the mid 1990s to gauge more accurately the amount of human activity on web servers. These were page views and visits (or sessions). A page view was defined as a request made to the web server for a page, as opposed to a graphic, while a visit was defined as a sequence of requests from a uniquely identified client that expired after a certain amount of inactivity, usually 30 minutes. The page views and visits are still commonly displayed metrics, but are now considered rather unsophisticated measurements.

The emergence of search engine spiders and robots in the late 1990s, along with web proxies and dynamically assigned IP addresses for large companies and ISPs, made it more difficult to identify unique human visitors to a website. Log analyzers responded by tracking visits by cookies, and by ignoring requests from known spiders.

The extensive use of web caches also presented a problem for logfile analysis. If a person revisits a page, the second request will often be retrieved from the browser's cache, and so no request will be received by the web server. This means that the person's path through the site is lost. Caching can be defeated by configuring the web server, but this can result in degraded performance for the visitor to the website.

### Page tagging

Concerns about the accuracy of logfile analysis in the presence of caching, and the desire to be able to perform web analytics as an outsourced service, led to the second data collection method, page tagging or 'Web bugs'.

In the mid 1990s, Web counters were commonly seen — these were images included in a web page that showed the number of times the image had been requested, which was an estimate of the number of visits to that page. In the late

## Web Analytics

1990s this concept evolved to include a small invisible image instead of a visible one, and, by using JavaScript, to pass along with the image request certain information about the page and the visitor. This information can then be processed remotely by a web analytics company, and extensive statistics generated.

The web analytics service also manages the process of assigning a cookie to the user, which can uniquely identify them during their visit and in subsequent visits.

With the increasing popularity of Ajax-based solutions, an alternative to the use of an invisible image, is to implement a call back to the server from the rendered page. In this case, when the page is rendered on the web browser, a piece of Ajax code would call back to the server and pass information about the client that can then be aggregated by a web analytics company. This is in some ways flawed by browser restrictions on the servers which can be contacted with XMLHttpRequest objects.

### Logfile analysis vs page tagging

Both logfile analysis programs and page tagging solutions are readily available to companies that wish to perform web analytics. In many cases, the same web analytics company will offer both approaches. The question then arises of which method a company should choose. There are advantages and disadvantages to each approach.

#### Advantages of logfile analysis

The main advantages of logfile analysis over page tagging are as follows.

- a) The web server normally already produces logfiles, so the raw data is already available. To collect data via page tagging requires changes to the website.
- b) The web server reliably records every transaction it makes. Page tagging relies on the visitors' browsers co-operating, which a certain proportion may not do (for example, if JavaScript is disabled)
- c) The data is on the company's own servers, and is in a standard, rather than a proprietary, format. This makes it easy for a company to switch programs later, use several different programs, and analyze historical data with a new program. Page tagging solutions involve vendor lock-in.
- d) Logfiles contain information on visits from search engine spiders. Although these should not be reported as part of the human activity, it is important data for performing search engine optimization.
- e) Logfiles contain information on failed requests; page tagging only records an event if the page is successfully viewed.

#### Advantages of page tagging

The main advantages of page tagging over logfile analysis are as follows.

1. The JavaScript is automatically run every time the page is loaded. Thus there are fewer worries about caching.
2. It is easier to add additional information to the JavaScript, which can then be collected by the remote server. For example, information about the visitors' screen sizes, or the price of the goods they purchased, can be added in this way. With logfile analysis, information not normally collected by the web server can only be recorded by modifying the URL.
3. Page tagging can report on events which do not involve a request to the web server, such as interactions within Flash movies.
4. The page tagging service manages the process of assigning cookies to visitors; with logfile analysis, the server has to be configured to do this.
5. Page tagging is available to companies who do not run their own web servers.

### Economic factors

## Web Analytics

Logfile analysis is almost always performed in-house. Page tagging can be performed in-house, but it is more often provided as a third-party service. The economic difference between these two models can also be a consideration for a company deciding which to purchase.

- Logfile analysis typically involves a one-off software purchase; however, some vendors are introducing maximum annual page views with additional costs to process additional information.
- Page tagging most often involves a monthly fee, although some vendors offer installable page tagging solutions with no additional page view costs.

Which solution is cheaper often depends on the amount of technical expertise within the company, the vendor chosen, the amount of activity seen on the web sites, the depth and type of information sought, and the number of distinct web sites needing statistics.

### Hybrid methods

Some companies are now producing programs which collect data through both logfiles and page tagging. By using a hybrid method, they aim to produce more accurate statistics than either method on its own. The first Hybrid solution was produced in 1998 by Rufus Evison who then spun the product out to create a company based upon the increased accuracy of hybrid methods.

### Other methods

Other methods of data collection have been used, but are not currently widely deployed. These include integrating the web analytics program into the web server, and collecting data by sniffing the network traffic passing between the web server and the outside world. Packet Sniffing is used by some of the largest e-commerce sites because it involves no changes to the site or servers and cannot compromise operation. It also provides better data in real-time or in log file format and it is easy to feed data warehouses and join the data with CRM, and enterprise data.

There is also another method of the page tagging analysis. Instead of getting the information from the user side, when he / she open the page, it's also possible to let the script work on the server side. Right before a page is sent to a user it then sends the data.

### Key definitions

There are no globally agreed definitions within web analytics as the industry bodies have been trying to agree definitions that are useful and definitive for some time. The main bodies who have had input in this area have been Jicwebs(Industry Committee for Web Standards)/ABCe (Auditing Bureau of Circulations electronic, UK and Europe), The WAA (Web Analytics Association, US) and to a lesser extent the IAB (Interactive Advertising Bureau). This does not prevent the following list from being a useful guide, suffering only slightly from ambiguity. Both the WAA and the ABCe provide more definitive lists for those who are declaring their statistics using the metrics defined by either.

- Hit - A request for a file from the web server. Available only in log analysis. The number of hits received by a website is frequently cited to assert its popularity, but this number is extremely misleading and dramatically over-estimates popularity. A single web-page typically consists of multiple (often dozens) of discrete files, each of which is counted as a hit as the page is downloaded, so the number of hits is really an arbitrary number more reflective of the complexity of individual pages on the website than the website's actual popularity. The total number of visitors or page views provides a more realistic and accurate assessment of popularity.
- Page View - A request for a file whose type is defined as a page in log analysis. An occurrence of the script being run in page tagging. In log analysis, a single page view may generate multiple hits as all the resources required to view the page (images, .js and .css files) are also requested from the web server.

## Web Analytics

- Visit / Session - A series of requests from the same uniquely identified client with a set timeout. A visit is expected to contain multiple hits (in log analysis) and page views. First Visit / First Session - A visit from a visitor who has not made any previous visits.
- Visitor / Unique Visitor / Unique User - The uniquely identified client generating requests on the web server (log analysis) or viewing pages (page tagging) within a defined time period (i.e. day, week or month). A Unique Visitor counts once within the timescale. A visitor can make multiple visits. The Unique User is now the only mandatory metric for an ABCe audit.
- Repeat Visitor - A visitor that has made at least one previous visit. The period between the last and current visit is called visitor recency and is measured in days.
- New Visitor - A visitor that has not made any previous visits. This definition creates a certain amount of confusion (see common confusions below), and is sometimes substituted with analysis of first visits.
- Impression - An impression is each time an advertisement loads on a user's screen. Anytime you see a banner that is an impression.
- Singletons - The number of visits where only a single page is viewed. While not a useful metric in and of itself the number of singletons is indicative of various forms of "Click Fraud" as well as being used to calculate bounce rate and in some cases to identify automatons ("bots").
- Bounce Rate / % Exit - The percentage of visits where the visitor enters and exits at the same page without visiting any other pages on the site in between.
- Visibility time - The time a single page (or a blog, Ad Banner...) is viewed.
- Session Duration - Average amount of time that visitors spend on the site each time they visit. This metric can be complicated by the fact that analytics programs can not measure the length of the final page view. Also, if a visit comes back to the site within a short period of time, that can be measured as a continuation of the first session.
- Page View Duration - Average amount of time that visitors spend on each page of the site. As with Session Duration, this metric is complicated by the fact that analytics programs can not measure the length of the final page view.
- Depth / Page Views per Session - Depth is the average number of page views a visitor consumes before ending their session. It is calculated by dividing total number of page views by total number of sessions and is also called Page Views per Session or PV/Session.
- Frequency / Session per Unique - Frequency measures how often visitors come to a website. It is calculated by dividing the total number of sessions (or visits) by the total number of unique visitors. Sometimes it is used to measure the loyalty of your audience.

### Common confusions in web analytics

#### The hotel problem

The hotel problem is generally the first problem encountered by a user of web analytics. The term was first coined by Rufus Evison explaining the problem at one of the Emetrics Summits and has now gained popularity as a simple expression of the problem and its resolution.

The problem is that the unique visitors for each day in a month do not add up to the same total as the unique visitors for that month. This appears to an inexperienced user to be a problem in whatever analytics software they are using. In fact it is a simple property of the metric definitions.

The way to picture the situation is by imagining a hotel. The hotel has two rooms (Room A and Room B).

	Day 1	Day 2	Day 3	Total
Room A	John	John	Jane	2 Unique Users
Room B	Mark	Jane	Mark	2 Unique Users
Total	2	2	2	?

## Web Analytics

As the table shows, the hotel has two unique users each day over three days. The sum of the totals with respect to the days is therefore six.

During the period each room has had two unique users. The sum of the totals with respect to the rooms is therefore four.

In actual fact only three visitors have been in the hotel over this period. The problem is that a person who stays in a room for two nights will get counted twice if you count them once on each day, but is only counted once if you are looking at the total for the period. Any software for web analytics will sum these correctly for whatever time period, thus leading to the problem when a user tries to compare the totals.

New visitors + repeat visitors unequal to total visitors

Another common misconception in web analytics is that the sum of the new visitors and the repeat visitors ought to be the total number of visitors. Again this becomes clear if the visitors are viewed as individuals on a small scale, but still causes a large number of complaints that analytics software cannot be working because of a failure to understand the metrics.

Here the culprit is the metric of a new visitor. There is really no such thing as a new visitor when you are considering a web site from an ongoing perspective. If a visitor makes their first visit on a given day and then returns to the web site on the same day they are both a new visitor and a repeat visitor for that day. So if we look at them as an individual which are they? The answer has to be both, so the definition of the metric is at fault.

A new visitor is not an individual; it is a fact of the web measurement. For this reason it is easiest to conceptualize the same facet as a first visit (or first session). This resolves the conflict and so removes the confusion. Nobody expects the number of first visits to add to the number of repeat visitors to give the total number of visitors. The metric will have the same number as the new visitors, but it is clearer that it will not add in this fashion.

On the day in question there was a first visit made by our chosen individual. There was also a repeat visit made by the same individual. The number of first visits and the number of repeat visits will add up to the total number of visits for that day.

### Web analytics methods

#### Problems with cookies

Historically, vendors of page-tagging analytics solutions have used third-party cookies that are cookies sent from the vendor's domain instead of the domain of the website being browsed. Third-party cookies can handle visitors who cross multiple unrelated domains within the company's site, since the cookie is always handled by the vendor's servers.

However, third-party cookies in principle allow tracking an individual user across the sites of different companies, allowing the analytics vendor to collate the user's activity on sites where he provided personal information with his activity on other sites where he thought he was anonymous. Although web analytics companies deny doing this, other companies such as companies supplying banner ads have done so. Privacy concerns about cookies have therefore led a noticeable minority of users to block or delete third-party cookies. In 2005, some reports showed that about 28% of Internet users blocked third-party cookies and 22% deleted them at least once a month.

Most vendors of page tagging solutions have now moved to provide at least the option of using first-party cookies (cookies assigned from the client sub domain).

## Web Analytics

Another problem is cookie deletion. When web analytics depend on cookies to identify unique visitors, the statistics are dependent on a persistent cookie to hold a unique visitor ID. When users delete cookies, they usually delete both first- and third-party cookies. If this is done between interactions with the site, the user will appear as a first-time visitor at their next interaction point. Without a persistent and unique visitor id, conversions, click-stream analysis, and other metrics dependent on the activities of a unique visitor over time, cannot be accurate.

Cookies are used because IP addresses are not always unique to users and may be shared by large groups or proxies. Other methods of uniquely identifying a user are technically challenging and would limit the trackable audience or would be considered suspicious. Cookies are the selected option because they reach the lowest common denominator without using technologies regarded as spy ware.

### Unique landing pages vs referrals for campaign tracking

Tracking the amount of activity generated through advertising relationships with external web sites through the referrals reports available in most web analytics packages is significantly less accurate than using unique landing pages.

Referring URLs are an unreliable source of information for the following reasons:

- They may or may not be provided by the web browser.
- They may or may not be recorded by the web server.
- They can be obfuscated intentionally by web browsers that wish to browse anonymously.
- They can be distorted or hidden by redirects, intentionally or not.