

WebTrends SmartSource Data Collection - Premier Client-side Data Collection Technology

Overview

More and more organizations are implementing an alternative technique to collect web site traffic information rather than relying on web server log files. This technique is called client-side data collection, or data tagging for short. Popularized by WebTrends and other web analytics vendors that provide hosted services, data tagging solves many problems associated with web server log file analysis.

With data tagging, web traffic data is more accurate because traffic normally hidden by cache or proxy servers is tracked. IT administration is eased because data collection is centralized in one location versus site data being dispersed among several log files from multiple web servers that may also be geographically dispersed. And web data can be collected from specialized applications, such as application servers and browser applications (e.g. Macromedia Flash).

With all these benefits of client-side collection, there are a few drawbacks. Implementing data tagging requires some development work to ensure that data tags are inserted and maintained on web pages. We'll discuss this process as part of this document. However, as you'll see, the benefits of client-side tagging far outweigh its drawbacks.

This paper will explain why client-side tagging is the best data collection method for analyzing web visitor behavior, and describe how it and log file analysis work. We'll also discuss why organizations should still analyze web server log files as a compliment to data tagging. Finally, we'll discuss our own client-side data collection technology, WebTrends SmartSource Data Collection, and explain why it is superior to other data tagging technologies.

Contents

The Challenges with Log File Analysis	2
How Data Tagging Works	3
Some Drawbacks with Data Tagging	4
Unique Advantages of WebTrends SmartSource Data Collection	5
Implementing Data Tagging	5
Summary	6

The Challenges with Log File Analysis

Web server log files, such as those produced by Internet Information Server (IIS), iPlanet or Apache, are detailed accountings of hits made against a web site. A hit is an individual request made to a web server from a browser. Intuitively one would think that a hit corresponds to a page on your site, but in fact most pages contain numerous objects, such as GIFs or JPGs, each of which produces a hit in the log file. It is therefore common for a single page request, or page view, to result in numerous hits being recorded in the log file.

So the first problem with web server log file analysis is that the vast majority of the data captured in log files has limited use in understanding visitor behavior. Many companies choose to filter out these superfluous hits from the analysis process, but the analysis engine must still read and parse each of these hits before determining if they can be disregarded, which of course adds processing time to the analysis process.

A related but more serious issue with web server log files is the potential accuracy problems they introduce. In many cases, web server log files do not accurately represent the actual interactions visitors have with a web site. Proxy servers are one of several examples of how analysis results can be distorted by web server log file data collection. Proxy servers deflect page views from web servers by caching the most frequently requested pages. Local caches have a similar effect, handling browser requests through locally cached pages rather than making repeated requests to the web server. In doing so, these page views are not recorded in the web server log files. Finally, many web sites are delivered by a Content Delivery Network (CDN) that distributes site content to potentially thousands of content servers worldwide. For web server log files, CDNs cause the same problem as proxy servers.

Due to these services, the most popular pages of your site may have a disproportionately small number of hits appearing in the log files. Likewise, path analysis reports could be of questionable accuracy, as an unknown number of page views may be missing from each visit. It's even possible for entire visits to be missing from web server log files if every page the visitor requested was delivered by cache, proxy or CDN, and the pages did not contain any non-cacheable elements. Other accuracy problems are created by spiders, crawlers, robots and other automated tools that scan your web site. These programs produce hits within web server log files, which can appear to be real visitors if their activity isn't filtered out.

Each of these issues means log files may not accurately reflect visitor behavior, and this can be a serious matter for organizations that need a precise understanding of how visitors interact with their site.

The most commonly observed problem with log files is the daily administrative burden they create. Collecting and analyzing log files from multiple web servers, especially if they are geographically dispersed, is often tedious and sometimes problematic. The collection of superfluous hits discussed earlier aggravates this problem, especially for large log files that need to be transferred via FTP.

One final problem is specific to Macromedia Flash-based applications. Flash applications do not produce log files. Web server logs will contain hits to pages containing the Flash application along with hits to the application itself, but no information is captured on how visitors behave within the Flash application. This is a rather profound problem for organizations choosing to use Flash. On one hand, Flash promises a far richer user experience than HTML-based pages, but on the other hand it's not possible to analyze the user experience using web server log files.

Data tagging solves all of the above problems while introducing a few of its own. If none of the aforementioned issues have a substantial impact on your business, feel fortunate and continue to use your log file analysis solution as you do today; otherwise, read on.

How Data Tagging Works

Data tagging, as the name implies, works by embedding tags on your web pages to transmit relevant data to a centralized data collection facility, or, in the case of WebTrends, the SmartSource Data Collector (SDC). The data is then analyzed directly from this central repository. The data tag itself is a small piece of scripting code, typically JavaScript, which transmits page-specific information via query string parameters. While the specific implementation varies between web analytics vendors, the general approach is the same across all commercial products and services that utilize data tagging.

The process begins when a page containing a data tag is requested from the web server. The tag is a piece of scripting code (either JavaScript or VBScript in SmartSource) that is executed when the page is loaded into the browser. The primary purpose of the SmartSource tag is to construct a GET request to the SDC for a 1x1 invisible GIF file. In addition to the name of the GIF file, the GET request contains query parameters loaded with page-specific data. The GIF file actually serves no purpose other than to provide a vehicle for transmitting the query string.

The SmartSource tag constructs the query string by scanning the document source for HTML META tags beginning with a specific identifier (“WT” in the case of SmartSource), such as the following:

```
<META name="WT.mc_n" content="Executive Mailer">
```

This META tag identifies the page as the landing page for the “Executive Mailer” marketing campaign through the predefined WebTrends query parameter “WT.mc_n.” There are many other predefined WebTrends query parameters, including content groups, ad views, ad clicks, scenarios, shopping cart contents, product names, product revenue and more. And of course developers can add their own parameters. Note that SmartSource is one of the few data tagging technologies that segregates page-specific information from the main script, maximizing code modularity and reuse for easier deployment and maintenance.

Once the document is completely scanned, a GET request is constructed by the SmartSource tag as follows:

```
http://SDCHostName/dcs.gif?dcsuri=/Path/URL.HTML&WT.mc_n=Executive%20Mailer
```

Upon receiving the request the SDC logs the hit into a SmartSource file, a standard W3C file. This is what allows WebTrends to offer data tagging in both its *On Demand* service and its software solutions. There are many important differences between standard web server log files and the data file that SmartSource creates.

First of all, the SmartSource tag is executed with each page view. This means the SmartSource file contains only one hit for each page view, regardless of how many objects are on the page. The result is that SmartSource files are much smaller than the corresponding web server logs.

Another substantial benefit is that the SDC produces a single centralized SmartSource file rather than separate log files for each web server. This essentially eliminates the administrative headaches associated with gathering logs from multiple, geographically dispersed web servers. The SmartSource file can even contain hits from multiple domains (the domain name can also be passed as a query parameter), allowing visitor behavior to be analyzed across an organization’s sites or even partner sites provided they permit your tags to be included on their pages.

Taken a step further, data tagging can provide information that is difficult or impossible to obtain with log files. For example, data tags linked to your SDC can be included in your banner ads placed on other sites. SmartSource tags can also be inserted into Flash applications, permitting a hit to be entered into the SmartSource file for each event fired in the program. This means visitor activity within Flash applications can be analyzed just like visitor interactions with HTML-based pages.

¹SmartSource supports dozens of pre-built query parameters.

One final benefit of data tagging is that it facilitates rapid availability of data for analysis. Because the data collection is centralized, analysis of the SmartSource file can occur as often as is necessary or practical, unlike web server log files that commonly need to be transferred on a nightly basis to the analysis machines. The ability to perform near real-time analysis and the use of data tagging itself, are often erroneously equated with hosted, on-demand services only, when, in fact, both benefits are also applicable to WebTrends software solutions. Both WebTrends 7 software and WebTrends 7 *On Demand* solutions support data tagging, allowing organizations to choose the delivery method that is best, without regard to the data collection methodology utilized. To-date, WebTrends is the only vendor to provide both software and *On Demand* solutions, and the only vendor to support data tagging in its software products.

Some Drawbacks with Data Tagging

While data tagging eliminates most of the administrative overhead of web server log files and captures more accurate visitor behavior analysis and information that was otherwise impossible to obtain, there are some drawbacks. The first and most obvious is that data tagging requires development effort. In order for data tagging to work, a piece of script must be placed on each page you want analyzed. In many cases this can be accomplished quickly by putting the script in a template, but there are also some instances where the script must be customized page to page.

Another problem is that data tagging doesn't capture some information that can be derived from web server log files. Most of this information is related to the hit-level diagnostics that web servers capture, including the amount of data transferred and web server load balance information. While this information is not needed to analyze visitor behavior, it is vital to IT staff responsible for optimizing web server availability and performance.

Data tagging also does not capture information related to spiders, crawlers, robots and other automated mechanisms. This data is useful in evaluating how deeply and frequently search engines are indexing your site, and in identifying which products are being scanned by price comparison services. If you own WebTrends software you can analyze your web server log files periodically to obtain this information, and use data tagging for analysis of visitor behavior. Note that only WebTrends provides both data tagging and web server log file analysis methodologies available in a single solution.

Data tags require additional bandwidth and, as a result, add to the wait time experienced by visitors as pages download. However, in most cases the added data transmissions for the data tags are small (in SmartSource, the script is less than 2KB and the GIF file is only 43 bytes).

For some organizations, the most problematic issue is that data tagging utilizes scripting and cookies (note that WebTrends SmartSource can be implemented without cookies). For most companies this is not an issue, but in some cases, such as government agencies, script is not permitted. Organizations with a strict anti-script policy cannot utilize data tagging.

One function of SmartSource not described earlier is the use of a persistent cookie sent by the SDC to the browser (in the default case). In SmartSource this cookie contains a single ID field used to uniquely identify a visitor. This way subsequent visits from the same visitor can be identified and used to determine the lifetime value of the visitor, the initial campaign that referred them to your site, and much more. Of all the vendors that provide client-side data collection, WebTrends is the only vendor that does not require the use of a cookie—this is one of the unique benefits of SmartSource Data Collection that we'll discuss next.

Unique Advantages of WebTrends SmartSource Data Collection

The most common problems organizations face in implementing data tagging are the use of cookies, due to privacy concerns, and the development resources needed to insert, test and maintain the tags. With WebTrends SmartSource Data Collection both of these issues are largely neutralized.

As mentioned earlier, cookies are not required in WebTrends SmartSource. But without cookies, information on returning visitors cannot be obtained unless you implement your own visitor identification method, such as authenticated logins. Besides identifying returning visitors, many web analytics vendors use cookies to store information about past visits, such as the date/time of the last visit, past purchase behavior and more. Thus disabling the cookie, even if it were possible with other vendor solutions, would render many marketing reports worthless. But in both WebTrends software and *On Demand* solutions historical visitor information is stored in a server-side table, called the Visitor History Table. So even if cookies are disabled in the visitor's browser, historical information will still be captured, using authenticated login as one example. The only thing stored in the WebTrends SmartSource cookie is an ID field used to identify the returning visitor, which can be supplanted by an alternative visitor identification technique.

If cookies are used, WebTrends SmartSource supports an unprecedented set of cookie configuration options. The software implementation of SmartSource permits the cookie to be first-party, as opposed to third-party. A first-party cookie is one that is served by your own domain, which eliminates some security concerns and reduces the number of cookies that are automatically blocked if visitors select medium to high privacy settings available in new web browsers such as Microsoft® Internet Explorer 6.0. Third-party cookies are served by an outside domain, the web analytics vendor's data center, which increases privacy and security concerns if cookies contain more information than a basic ID field. When evaluating a web analytics on-demand service, it's critical to understand what information is maintained in the cookie.

The other primary concern with data tagging is the development effort required to insert, test and maintain the tags on your web site. Most web analytics on-demand vendors merely provide a data tag where all of the page-specific information is within the JavaScript itself. The more customized the data you want to collect, the more effort will be required to modify and insert scripts page to page. To simplify this process, the WebTrends SmartSource tag contains page-generic code in the JavaScript and allows page-specific information to be placed in the META tags. We'll discuss the benefits of this approach, along with the general implementation process next.

Implementing Data Tagging

Incorporating data tags into a site is not much different than inserting any other code. As explained earlier, there are two parts to the SmartSource tag: the first is the script that constructs the GET request to the SDC, and the second is a series of META tags that identify the analysis settings that are specific to the page.

In general, there are four steps to the process of implementing data tags on a site:

1. Identify the pages that require data tagging and determine the methodology for inserting the script on each page
2. Determine the analysis settings for each page
3. Insert the scripts and meta tags onto your pages
4. Test and debug the tags before checking the code on the staged or live site

The process detailed above largely assumes you are retrofitting an existing site with data tags. But that does not necessarily need to be the case. In fact, the greatest potential benefit to data tagging is that it encourages site designers and developers to think about web analytics in the early stages of the site design process. Organizations that incorporate web analytics into their design work are much more likely to create better user experiences and more effective sites.

There are several approaches to inserting the script into the pages of your site. The most obvious is to insert the entire script into each individual page. But a better approach is to embed the script into a template, allowing changes to the script to be made globally (the script can be also be incorporated into a #include file through an <include> tag or through a server-side include). Inserting scripts using templates or #include files is only practical if the script contains page-generic code.

Once development is completed, the SmartSource tags can be used with either WebTrends software or *On Demand* solutions by simply changing the domain name of the SDC from the local test copy to the production server or service. One of the most compelling advantages of SmartSource is that you can easily migrate your web analytics solution from WebTrends *On Demand* to WebTrends software, or vice versa, with virtually no change to the SmartSource tags.

Summary

Client-side data collection is quickly growing in popularity as the superior approach to collecting web visitor behavior information. It provides greater reporting accuracy and lower administrative overhead. Organizations should carefully analyze the costs and benefits of data tagging versus web server log file analysis, and determine which method will best meet your insight needs.

Organizations also need to thoroughly evaluate the data tagging technologies of each web analytics vendor they are considering, as some have security and privacy issues, while others do not help with the implementation of the data tags. WebTrends SmartSource Data Collection is clearly superior to other data tagging technologies available:

- SmartSource is the only client-side data collection technology available as both software and an *On Demand* service, permitting organizations to freely migrate between software and service, or to incorporate a combination of the two.
- SmartSource offers the greatest level of security and privacy protection with the most flexible cookie configuration available.

SmartSource is one of many ways WebTrends provides its customers with the flexibility and choice they need to improve their web sites with more effective web analytics. Most organizations recognize that their analytical needs will change over time; therefore, it's critical to choose a web analytics vendor and solution that will grow with you and adapt as your needs change.