VisiStat
Where Clicks Count.™

HOSTED WEB ANALYTICS VERSUS SERVER LOG FILES

As the demand for meaningful analytics increases, it has prompted new thinking about the means of collecting data for the purposes of analysis. For many years, Web server logs, also known as clickstream data, have served as the definitive source of online data. Ease of availability explains much of their popularity. Web servers were configured to produce a record, a log, of activity on the site. Logs record such things as error and status messages, requests for pages and transaction details; they were designed to serve the needs of site administrators, but were re-purposed by marketing researchers eager for new sources of information. Increasingly sophisticated demands for analysis of business metrics, as opposed to site metrics, have revealed shortcomings in log files. Consider the following quote from the usability-testing firm, *User Interface Engineering*, describing recent experiments with log files:

"The term 'data swamp' can be appropriately applied to Web logs. There's a lot of data to trudge through before you find the information that can really be of use. We've found that Web logs can function as valuable tools for programmers but aren't really designed to provide information pertinent to a company's bottom line."

Source: User Interface Engineering, Eye for Design, February, 2001

We feel that "snippet" technology offers superior data-collection and analysis capabilities when compared to log files. The superiority of snippet technology to log files rests on several key points:

- snippets produce much more accurate data than log files;
- information collected using snippets has smaller storage requirements; and,
- snippet technology facilitates real-time reporting as opposed to batch-file processing required for log files.

The following paper describes in greater detail how snippet tracking and log files differ in terms of how they collect data and what they collect. The paper also explores issue of data accuracy and data storage and handling.

HOSTED ANALYTICS

As the demand for meaningful analytics increases, it has prompted new thinking about the code snippets. When read by a visitor's browser, several small JavaScript executables are triggered, which run on the visitor's computer and collect visit information. Information collected is non-identifying and includes user behaviors such as visits and buys, responding to off-site or on-site promotions, and product viewing. Snippets also permit the collection of user characteristics, including profiles of their hardware and software used to view the site. When combined with cookies, tiny text files stored on visitor computers, snippets are able to provide information on pages visited, number of visits, and various timing and duration measures.

LOG FILES

In contrast, log files collect records of server requests made by visitor browsers. When a visitor comes to a site, each Web page, graphic, and some other components are separately requested from the host server and logged as individual requests. It is these requests that make their way into the log files, along with some information about time and other basic factors. Log-file analysis will most often take the accumulated log files for a set period of time and process them as a batch.

ACCURACY OF DATA

The opportunity to track and measure everything makes the Internet the marketer's best friend. Yet, the value of any research is questionable if it is based on a poor methodology and less than accurate data. There are clear signs that log files, which were designed as an administration tool before being eagerly re-purposed by marketing researchers, do not provide the necessary accuracy. Log files have been estimated to miscount visits by as much as 40%.

CACHING

Another feature of most browsers is the ability to cache visited Web pages. Caching is when Web pages are accessed and then stored in another location, either a users' hard drive or a third party server. Caching eliminates the need to re-download pages from the Web site server when it is subsequently requested by a visitor. Server log files only record requests made for a page. If a page is cached, the request will not reach the Web server logs and will therefore not be recorded. The prevalence of any of these caching methods is likely to vary widely depending upon the popularity of a site. Popular sites are more likely to have pages cached, and therefore, have proportionately more pages uncounted in log files.

VisiStat's "snippet" has the advantage of being embedded within the page itself, whether accessed directly from the host Web site, or from a cache. Thus, the data tag will "fire" (allowing the recording of a visit and other pertinent information) when the Web page it is on is accessed, whether or not it is accessed from a cache or the host server. Thus, in this regard, counts of page looks or visits based on data tags will be more accurate than log-file analysis.

SPIDERS AND BOTS

Spiders and bots are terms for software programs that "crawl" across the World Wide Web, usually to archive and classify sites and site content for various search engines. Many of the major search tools, and likely many more minor or local engines, use these programs to create search indexes and keyword databases. These types of tools have also become popular among online shoppers to automate price comparisons among Web site vendors. The important thing about these programs is that they often work by making requests of the server to access Web page information, and these requests appear in the server logs as legitimate requests, which will then be reported by log-file analysis tools as legitimate visits and page views. In contrast, most spiders and bots do not "fire" data tags (i.e., execute the JavaScript) that are imbedded in pages, preventing this method from erroneously inflating counts of visits and visitors.

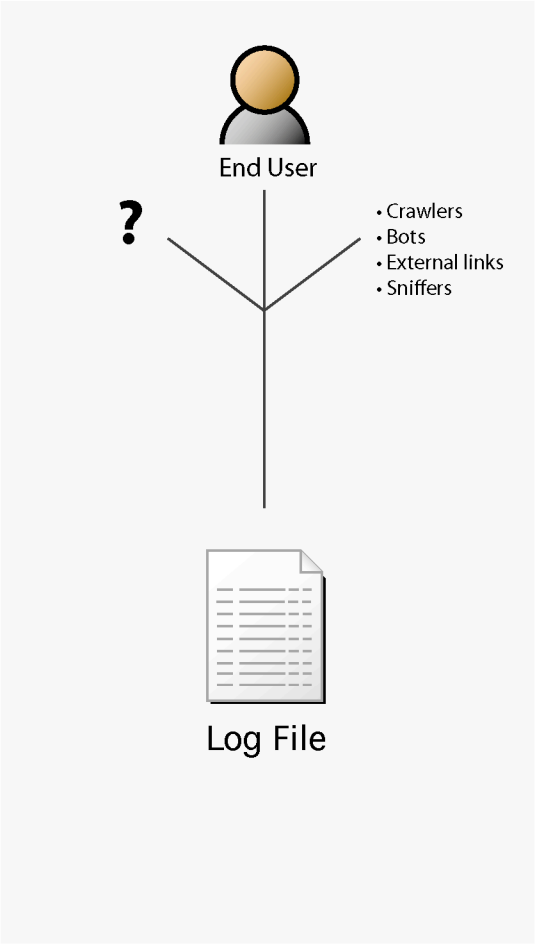
PERFORMANCE AND DATA STORAGE

Any discussion of the relative merits of hosted analytics vs. log files has to address back-end type issues such as performance and data storage. Log files have performance and collection costs. The collection and storage of log files demand processing cycles and memory from Web servers. For large sites with millions of hits a day, the cost of these performance decrements can be considerable. The costs are compounded in the case of large Web sites with multiple-server configurations, which create numerous silos of log data that must be aggregated and cleaned before analysis can take place. And, one should not forget the effort and resources necessary to analyze this data, which, of course, does not take place in real time. This fact alone raises questions about the accuracy and usefulness of log-file data; for time-sensitive factors such as the launch of new products and campaigns, it seems unreasonable to have to wait to receive reports on "old" data.

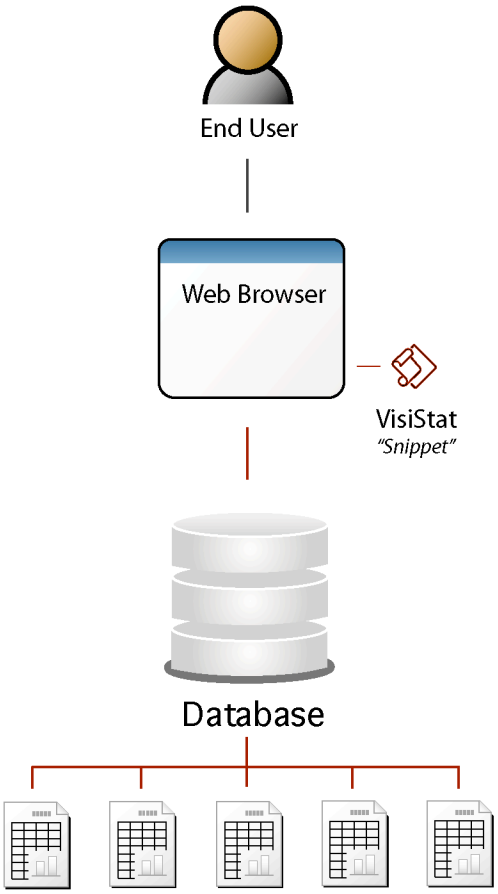
Hosted analytics eliminate many of the problems associated with storage and performance. To start, the VisiStat hosted analytics solution eliminates the need to store, organize and prep log files. In addition, because VisiStat captures the data in real-time, the processing power to capture and organize user data is performed on dedicated servers. The most significant advantage of this system facilitates real-time reporting. In other words, business analysts can look at the data whenever they want and see up-to-the minute metrics. This is a major consideration when dealing with time-sensitive factors such as advertising campaigns, the launch of new promotions, etc., all of which require quick feedback for forecasting purposes.

DATA GATHERING AND STORAGE MODEL

Log File Method



VisiStat "Snippet" Method



CONCLUSION

In this paper we have discussed the merits of log files relative to hosted analytics, identifying a number of important areas where VisiStat's snippets solve problems encountered with log files. Accuracy is paramount when dealing with data collection and analysis. The accuracy of data harvested via log files, however, is endangered by a number of factors such as caching, proxy servers and spiders/bots. Using the VisiStat approach, which sits at the page level, one avoids each of these pitfalls. Page snippets represent a strategic method of data collection, a fact that reduces storage requirements, facilitates real-time reporting and ultimately delivers more useful data than can be derived from log files.