WEB ANALYTICS: Client-side and server-side methods of data collection and analysis

MUAZ CISSE, MARIAM DEMBELE

Research work in Information Systems / Marketing

Supervised by:
Prof. Dr. Bambauer-Sachse
Prof. Dr. Andreas Meier

Fribourg, May 2009
# Table of contents

List of tables .......................................................................................................................... iv
List of code ................................................................................................................................ v

1 Introduction ........................................................................................................................... 6

2 Data Collection and Analysis .............................................................................................. 8
   2.1 Why data collection? ........................................................................................................ 8
   2.2 Why analysis? .................................................................................................................. 9

3 The Internet, a stateless world ............................................................................................ 10

4 A qualitative approach of the Customer Relationship Management: web analytics .......... 12

5 The public records of the server ......................................................................................... 15

6 Listening in the network ...................................................................................................... 17

7 Tag along visitors ................................................................................................................ 19

8 An efficient monitoring recipe ............................................................................................ 20

9 Relying on the server .......................................................................................................... 23

10 Banking on a remote ally ................................................................................................... 25

11 Case study .......................................................................................................................... 27
   11.1 Overview of the case study .......................................................................................... 27
      11.1.1 Self-hosted analytic solutions .............................................................................. 28
      11.1.2 Third-party hosted analytic solutions ............................................................... 38
   11.2 Key Findings ................................................................................................................ 41

12 Conclusion .......................................................................................................................... 43

Bibliography ............................................................................................................................ 45
List of figures

Figure 1: The logic of continuous improvement [Massot, 1999] ........................................ 12
Figure 2: Customer Life Cycle Funnel [Kaushik and Waisberg, 2009] .............................. 14
Figure 3: The web logs ........................................................................................................ 16
Figure 4: The network analysis ............................................................................................ 17
Figure 5: The page tagging technique ................................................................................ 19
Figure 6: The ingredients of a cookie .................................................................................. 21
Figure 7: BBClone admin interface .................................................................................... 28
Figure 8: ClickHeat admin interface .................................................................................. 29
Figure 9: FireStats admin interface .................................................................................... 31
Figure 10: Grape admin interface ....................................................................................... 32
Figure 11: phpMyVisites admin interface ......................................................................... 33
Figure 12: Piwik admin interface ....................................................................................... 35
Figure 13: SlimStat admin interface .................................................................................. 37
Figure 14: Clicky admin interface ...................................................................................... 38
Figure 15: Google Analytics admin interface .................................................................... 40
List of tables

Tableau 1: Client-side vs. server-side .......................................................... 11

Tableau 2: Web analytics solutions benchmark ........................................... 41
<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code 1</td>
<td>Code customer identifier</td>
<td>20</td>
</tr>
<tr>
<td>Code 2</td>
<td>Code web pages the customer visited</td>
<td>20</td>
</tr>
<tr>
<td>Code 3</td>
<td>Code items the customer has in his/her basket</td>
<td>20</td>
</tr>
<tr>
<td>Code 4</td>
<td>Code orders the customer placed</td>
<td>21</td>
</tr>
<tr>
<td>Code 5</td>
<td>Activation code source for BBClone</td>
<td>29</td>
</tr>
<tr>
<td>Code 6</td>
<td>Activation code source for ClickHeat</td>
<td>30</td>
</tr>
<tr>
<td>Code 7</td>
<td>Activation code source for FireStats</td>
<td>32</td>
</tr>
<tr>
<td>Code 8</td>
<td>Activation code source for Grape</td>
<td>33</td>
</tr>
<tr>
<td>Code 9</td>
<td>Activation code source for phpMyVisites</td>
<td>34</td>
</tr>
<tr>
<td>Code 10</td>
<td>Activation code source for Piwik</td>
<td>36</td>
</tr>
<tr>
<td>Code 11</td>
<td>Activation code source for SlimStat</td>
<td>38</td>
</tr>
<tr>
<td>Code 12</td>
<td>Activation code source for Clicky</td>
<td>39</td>
</tr>
<tr>
<td>Code 13</td>
<td>Activation code source for Google Analytics</td>
<td>40</td>
</tr>
</tbody>
</table>
1

Introduction

Last year, 8000 challengers from 60 countries were competing for the ‘Webby Awards’ better known as the ‘Internet Awards’ like the New York Times dare say. Organized by the International Academy of Digital Arts and Sciences, this prize rewards ‘Interactive Advertising, Online Film & Video’ and contents for Mobile devices. [WebbyAwards, 2009] They especially recompense websites from 30 categories\(^1\) according to the following criteria: ‘Content, Structure and Navigation, Visual Design, Functionality, Interactivity, Overall Experience’. As the Wall Street Journal put it into words, the Webby Awards “[celebrates] sites that pave important paths to the Internet’s next phase.” [WebbyAwards, 2009]

This last statement underlines the fact that we are in a new generation of the virtual world, where people do not have web sites only for the sake of it. Every e-business webmaster has two special tasks. The first one is to build a unique web site, capable of attracting individuals due to both its content and graphics. The second one is to enable a company to sell its products online and make high profits. Nevertheless “web development, is not an exact science” [WebbyAwards, 2009]. There is no miracle cure that can ensure that you produce a fine web page and sell the products of your company. Yet it is possible to rely on a single person that can give you an honest feedback on your work, the visitor.

Internet users can visit your website at any hour from different locations around the world, each of them with different motives or needs. Some may register for an account, leave a comment, send you an email, browse a product’s information or buy it, etc. [Dosh Dosh, 2009]

Unlike physical retail stores, where you can see who is coming in and browsing around, you do not know much about the customers’ habit. How can a rough profile of all these individuals visiting the website be developed, especially when we know that Internet users are not very keen on answering to online questionnaires?

\(^{\text{1}}\) The Webby Awards honors for example web sites from the following sectors of activity: Activism, Art, Associations, Automotive, Banking/Bill Paying, Beauty and Cosmetics, Blog - Business, Celebrity/Fan, Charitable Organizations / Non-Profit, Consumer Electronics, Cultural Institutions, Education...
Introduction

To answer this question, we can turn to techniques which are derived from the combination of visible user actions (e.g. comments, emails, surveys) and statistics (e.g. web analytics). This work will focus on Web analytics, which are tools that scrutinize the behavior of visitors on web pages and collect data about Internet users’ preferences. In this paper, we will study the miscellaneous available technologies used by webmasters, marketers and media savvy professionals to learn even more about their audiences; with the goal of providing answers to the following questions: What are the technical solutions conceived in order to fit as closely as possible the customer wishes? How do they work? What are their advantages and drawbacks? How many visitors came to our site? How long did they stay? How many pages did they visit? How to better match a customer’s needs to the right analytics package?

Our answers to these substantial questions will be presented in two main parts. At first, we will give reasons for using web analytics. Then we are going to examine the existing tools from a theoretical point of view. We will analyze server, network and client-side applications. Thereafter we will adopt a more practical approach, giving concrete use cases of web-based analytical tool.

There are 9 analytical tools examined in this paper. They are: BBClone, ClickHeat, Clicky, FireStats, Google Analytics, Grape, Phpmyvisites, Piwik and Slimstat. The use case section will help us understand the basic capabilities of each package, look at the ease of use of each analytical solution, how they can be configured to better match customer’s need to the right analytical solution, and solve specific problems on each web site.
Data Collection and Analysis

Data collection and analysis are helpful to collect, analyze and construct a general profile of our audience, however shifting it may be, because it provides us with information that will allow us to better improve our content scope, site usability, conversation rate or marketing campaign. Let’s split this process up into two sections: data collection and statistical analysis that we are going to explain.

2.1 Why data collection?

Companies in the web space spend millions of dollars to know how visitors interact with their websites. Consistently the number one challenge in surveys, case studies, and fix-it wish lists is the ability to measure accurately to make optimal decisions for the millions of dollars companies spend.

What motivates people to go through the often complex and costly process of collecting data? Apart from simply collecting information to satisfy a fascination with so-called trivia, most data are collected for a specific purpose, as part of a broader strategy. We may be surveying how people would react to the introduction of a new distribution strategy, or the likely sales for a new product. The following are some of the main reasons why people collect data:

Propaganda

Some data are collected for what we might call propaganda purposes: to convince other people of the rightness of your view, or a group to which you belong. Most propaganda that involves real data is based on processing and presenting raw data in a way that suits a particular message, rather than on the generation of new data.

Market research

Enormous amounts of data are collected by commercial organizations about the buying intentions of consumers; these surveys are also widely used in the social arenas.
Decision support

In industry and government, we have become used to the expectation that decisions will be based on careful analysis of data. For example, before introducing a new product to the market, a business analysis will be carried out to determine the impact of the new product to the existing products or the market. Relevant data about the products and market affected are collected and collated; this would then form one of the key foundations of the decision about whether or not to proceed with the introduction of the new product.

In the next section, we will examine why we need to explore data obtained from these tools.

2.2 Why analysis?

In marketing and advertising, we proactively define our target audience. We start with our end goals and then structure our website/ads with the right buzz phrases, pitch, style, keywords and angle to appeal to people we want to attract as a consumer/user/reader. Gathering information on visitors to our website makes us more effective marketers. [Dosh Dosh, 2009]

It is true that some researchers collect data but many do not correctly analyze the data collected. Software and hardware applications used to collect the data are in the vast majority of cases incapable of interpreting the data for us. This is due to the differing reasons for collecting the data. For example, an online support website will have as an objective to minimize the total time each visitor spends on their website whereas an online news website might have as an object to maximize the total time each visitor spend on their website. Also depending on the nature of the website, the definition of the total time spent on the website might vary. Other websites might assume that Time on Site = Time of engagement. This means that the time the visitor spent on the site interacting with it directly. [Kaushik, 2007]

Another reason for data analysis is that most web analytics tools differ in the way they define the different metrics used to capture data as there is no standardization in the industry. Thus data analysis will help the enterprise define the data capture methodology in a way that best fits their needs.

After questioning us about the role of the data collection and analysis, we are now going to see virtual solutions at our disposition to perform this task.
The Internet, a stateless world

The Internet communication is based on client-server architecture. As in all exchanges, two parties are involved. On one hand, a computer called ‘server’ offers valuable services and stores useful information. That is where the webmaster and the IT-specialist work to keep the information of a company reliable and up-to-date. On the other hand, a workstation named ‘client’ requests for data. It is usually the customer’s computer. This mode of interchange enables computers all over the world to communicate without being physically linked.

From an e-commerce point of view, the problem of this infrastructure is that the World Wide Web is stateless. During the communication, the client and its host share messages. Nevertheless as soon as the interaction is over, the server has no mechanisms to identify the workstation it had a dialog with. This leads to the fact that the company cannot recognize the customer it was offering its product to as well. Moreover misunderstandings might take place. A large number of visits can occurs when many customers visit an Internet page. Similarly the same results appear when a customer visits many times the same page.

At this point, we need to resort to web analytics to help us understand what happens on a website. These software programs scrutinize from the server or the client-side the behavior of the customer that navigates on our website. To explain the difference between server-side en client-side analytics, one first needs to understand the singularity of server-side and client-side scripting.

- ‘Client-side scripting’ means that all calculations are done by the computer of the visitor. Most of the time the script is interpreted by their internet browser (Javascript) or a specific browser plugin (Flash actionscript).

- ‘Server-side scripting’ takes place when all calculations are done by the server your website is hosted on. The script is decoded by the supported language parser, like PHP or ASP.

The following table shows information available in both types and the main restrictions. [Linklove, 2009]
<table>
<thead>
<tr>
<th>Client</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browser properties</td>
<td>Server/site/page properties</td>
</tr>
<tr>
<td>Page element properties</td>
<td>Connection properties</td>
</tr>
<tr>
<td>Change possible on any interaction</td>
<td>Change requires reload</td>
</tr>
<tr>
<td>Read-only access to external files</td>
<td>Many communication possibilities</td>
</tr>
<tr>
<td>Entire script source visible</td>
<td>Script source hidden for visitor</td>
</tr>
<tr>
<td>Restrictions based on browser security-and support settings.</td>
<td>Restricted by server settings</td>
</tr>
<tr>
<td>Client support required</td>
<td>All requests are counted</td>
</tr>
</tbody>
</table>

**Tableau 1:** *Client-side vs. server-side*
As utters Avinash Kaushik, a specialist of website inspection, “Web Analytics is the science and the art of improving websites to increase their profitability by improving the customer’s website experience. It is a science because it uses statistics, data mining techniques, and a methodological process” [Kaushik and Waisberg, 2009]. As a matter of fact, the website supervision is thus a matter of quality and aims to fulfill at best the needs of the customers using Customer Relationship Management techniques.

The first step of this willingness to improve the provided services is the use of the Deming cycle and in particular the “plan-do-check-act” modus operandi, represented in the figure 1.

Figure 1 : The logic of continuous improvement [Massot, 1999]
At first, one needs to define the goals to be reached, i.e. PLAN what has to be done. It follows that metrics, called ‘Performance Key Indicators’, have to be defined in order to guide the contributed efforts. They include the followings:

- the hits, that shows which Internet page were delivered by the server,
- the page views, that tells us how many time a web page had been asked for,
- The unique visitor or an individual that was surfing on one’s website.

From these, suitable other indicators can be determined:

- “percent of user growth this period,
- percent of traffic growth this period,
- percent of traffic from search engines,
- average page view per visit,
- average visit per visitor,
- percent low/medium/high time spent visit,
- percent low/medium/high click depth visit” [LeClaire and Sostre, 2007].

In the second phase, DO, these metrics will help delineating the measures to be applied to get successful commercial and strategic results. One aim is then to make every hit become a purchase and change a customer to a consumer, or at least secure the loyalty of the client. This latter will depend on the website’s interface and functionalities.

One has then to CHECK if the means implemented are really efficient, according to the most important stage of the continuous improvement cycle. That is when the web analytics come into play. Indeed thanks to them, one can obtain facts on the customer preferences with information on: the data traffic, the rush hours, the files transferred, the most popular webpage, the least visited resources…

The obtained material helps to ACT properly and deploy one’s energy in the right direction to establish a user-oriented development of a web site, which leads to a more efficient online marketing. It tightens client to our company. Moreover it helps avoiding the martini, margarita or wine glass shaped behavior of the client, where there are many acquisitions of customer but at the end only a few consumers and conversion rate, as shows the schema 2.
A qualitative approach of the Customer Relationship Management: web analytics

Figure 2: Customer Life Cycle Funnel [Kaushik and Waisberg, 2009]

The collection of information needed to establish a special relationship with the visitor, can be accomplished in two different ways. In the first one, web analytics gather server-side information with web logs.
Web analytical tools exist for quite a long time. They are artifact of the client-server architecture. At the beginning of the Internet, they were created to assist computer scientists examining the various actions happening on their network and debug the dysfunctional services if necessary. Web logs purpose was consequently to record every solicitation of a resource on the server. When a client computer calls a web page, the browser it uses makes a request to a host. This latter is saved as a ‘hit’ in a text document, the ‘log file’.

Later this technique was used by marketing specialists to track visitor’s journey on web pages, due to the various data it provides. In fact, visitors are identified by an IP address. This digital identity card is taking into consideration for each demand of the client machine. With this information the domain a user comes from, can be identified. One can thus know among other things from which country a visitor comes. Many other parameters also play an important role:

- the name used for the login,
- the date and the moment at which the page was requested,
- the URL of the demanded information,
- The type of resource asked for (Internet page, pdf file, picture…),
- the name of the browser and operating system used,
- the code the server gives as a response, for example 404 for an invalid link, 200 for a successful communication… This helps detecting the server-side bugs.
- the referrer or Internet page the visitor has already seen,
- The size of the document requested.

Below is a typical example of a log file record [Apache, 2009].
The public records of the server

At the end of the exchange, an answer can be given to the demand and the desired page is sent. When these web files are produced, a program, the server-side application, is called to process the information and examine the data obtained.

The log files are also used in another technique called the ‘network analyses’.

Figure 3: The web logs
The network examination, based on ‘packet sniffing’, was in the beginning like log files used by computer scientist and in particular network specialist to analyze traffic, protocols, and packets and prevent hacker’s intrusion on the network.

A packet ‘sniffing’ software is therefore, in this technique, installed between the enterprise network and Internet. It can be deployed on the company’s server or on a separate device; in this case the enterprise carries out a server farm architecture. In this later configuration, it can be a hub (a device that connects servers or computers so that they can share information on the local segment), or a router that couples Internet domains. In fact, the workstation is attentive to the Ethernet protocol used for the local communication of workstations. It captures every data packet sent with the protocols of the TCP/IP stack to or from one’s network. Information about the workstations involved in an exchange, the protocol used, and the time needed to send information from the source to the destination… are provided as in the figure 4 below. The acquired data is processed and generates a log file.

![Figure 4: The network analysis](image)

<table>
<thead>
<tr>
<th>No.</th>
<th>Time</th>
<th>Source</th>
<th>Destination</th>
<th>Protocol</th>
<th>Info</th>
</tr>
</thead>
</table>
| 1   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 2   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 3   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 4   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 5   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 6   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 7   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 8   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 9   | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 10  | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 11  | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network
| 12  | 0.00000 | 192.168.1.20 | 192.168.144.75 | TCP | Local network

Figure 4: The network analysis
The second method of gathering information about a visitor is to tag web pages.
The most appreciated method of web analytics is certainly ‘page tagging’. This technique functions as follows: a JavaScript code is inserted on a web page. As soon as the client opens the requested page, the browser loads the program. A second approach uses ‘web beacons’, a small image often translucent, of maximum one pixel, placed like JavaScript programs on a web page and loaded by the browser to follow a visitor’s itinerary on the web. The only difference with the JavaScript method is that the beacon often comes from a third party server.

Both instruments record the actions done by the customer on his/her computer. The numerical identifier of the page and the customer surf habits are sent to the web analytics server for further analysis. The information is processed remotely to produce easy-to-read graphs and charts as indicated the schema 5. Page tagging programs can be maintained internally or outsourced to an Application Service Provider to which a monthly fee has to be paid.

In the page tagging method, cookies, storing data about the visitor behavior in a domain, are attributed to the customers.
An efficient monitoring recipe

As we have seen, the communication between the client and its server is stateless on the Internet. We therefore need a mechanism that allows the server to remember the computer it was dealing with. That is when the cookies are involved.

As does not their name suggest, cookies are text documents saved on the client hard drive to make them recognizable for a given server. They are of two types: first party cookies coming from the server we are dealing with and third party cookies from a remote workstation. These files can store diverse type of information:

- the customer identifier,

```
Set-Cookie: Customer="WILE_E_COYOTE"; Version="1"; Path="/acme"
```

Code 1: Code customer identifier

- the web pages the customer visited,

```
POST /acme/pickitem HTTP/1.1
Cookie: $Version="1"; Customer="WILE_E_COYOTE"; $Path="/acme"
```

Code 2: Code web pages the customer visited

- the items the customer has in his/her basket,

```
Set-Cookie: Part_Number="Rocket_Launcher_0001"; Version="1"; Path="/acme"
```

Code 3: Code items the customer has in his/her basket

---

2 The method ‘set-cookie’ specifies the parameters of the cookie.

3 The procedure POST signals that the server sends a resource to the client.
the orders the customer placed,

```plaintext
POST /acme/process HTTP/1.1
Cookie: $Version="1";
Customer="WILE_E_COYOTE"; $Path="/acme";
Part_Number="Rocket_Launcher_0001"; $Path="/acme";
Shipping="FedEx"; $Path="/acme";
```

**Code 4 : Code orders the customer placed**

We can notice that the data collected by the cookie is only made of messages sent by the client computer to the server. The cookie has no mean to collect information from the customer hard drive and breach his/her privacy.

At the end of the day, this leads to the fact that the next time the customer will call the same Internet page, the server will check if a cookie under its authority is stored on the computer. If it is the case, customized news or offers will be sent to the client.

Cookies are made of 6 fields, like shows this example, taken from the Amazon web site.

![Cookies](image)

**Figure 6 : The ingredients of a cookie**

First, the compulsory field ‘name’ has to be declared. After that the content, formed by unique and arbitrary chain of character, is specified. The Internet domain it came from has to be given and verified for security reasons before the exchange takes place. This later is followed by the path of the document, i.e. the directory of the

---

4 A cookie is created for the different requests of the customer.
server in which one can find it. Here for example, the cookie is in the current directory of the server of Amazon. In addition, the type of connection used to send the cookie is identified. This gives an answer to the question: ‘does the server put every type of connection or only secured ones to use?’ Moreover the lifetime of the cookie is declared.

As a matter of fact, cookies utilization is limited to avoid misuses. As shown in the figure, after a certain deadline close to the duration of the communication session\(^5\), the cookie expires. The obtained information cannot be sent to the server anymore. The memory occupied is released for the server to record more actual information. The other limitations are of size and number, i.e. a cookie must not exceed 4KB (4096 bytes). Furthermore an Internet domain cannot send more than 20 cookies to a client machine, due the fact that a computer cannot store more than 300 cookies.

After explaining the different technologies used, we are going to discuss the advantages and the limitations of each of them, starting with the server-side solutions.

---

\(^5\) These cookies are called the ‘session cookies’. They are in opposition of the ‘persistent cookies’, which lifetime is longer than the session duration.
Server-side solutions have many advantages as:

- They are already installed and configured on the network. Servers produce log files automatically without increasing the bandwidth capacity.
- The acquired information tends to be more reliable, owing to the fact that the same device records the traffic and generates the log files.
- The information is processed in-house and not sent to a third party entity.
- The results do not depend on the willingness of a visitor to enable JavaScript on his/her browser, nor on the web analytical tool.
- Spiders and robots as taken into consideration, enabling the optimization of the search engine.
- Historical data can be accessed easily.
- Log files take failed requests into consideration, i.e. they track download errors.
- Log files are conceived to have a standard. There is no switching cost if one wants to change the program employed.

Furthermore log files record usernames, in case of misuses of the network. The network analysis has the same avails as the server-side method for it is also based on log files. Nevertheless it offers additional functionalities. One is centralizing the data of a server farm into a single document. Facts about network intrusions and equivocal content are also available to supervise properly the network.

These points turn out to be drawbacks of the client-side applications:

- To implement page tagging, one has often to see the service of an external ‘Application Service Provider’ that will charge a monthly fee.
- Statistics about the website traffic can be lost due to storage space or network problems for hosting.
Page tagging software programs are proprietary. If one has to switch, cost will be induced.

Due to the fact that the data is sent to a remote server and cookies are being involved, this method has some confidentiality and privacy issue.

Statistics previous to the page tagging technique are not taken into consideration. This information is on the contrary yours, as long as you did not erase it from your server.

User can be included twice in the statistics in case they changed their computers or the browser used.

If the visitor's browser does not enable JavaScript, lots of information can be lost.

Adding JavaScript to a page extends its size and slows the communication between two machines because of the bandwidth's space occupied.

You have to insert JavaScript tags on every page to be analyzed. It takes therefore a lot of time to be accomplished and sustained.

It is not possible to include a JavaScript tag in a downloadable file. One can only notice when the process started but not if it was effective or the contrary. Moreover solely successful viewed pages are considered. Errors are accordingly not reported.

Client-side applications present additional information invisible to log files.
Client-side programs also have benefits:

- An enterprise does not need to have its own server or a computer specialist to get web analytical statistics.
- The reports are obtained faster; data is processed in ‘real time’.
- The assignment of cookies to visitors is outsourced to the hosting company.
- Cookies help making a difference between group of users that are using an identical IP-address like colleagues, friend and family using the same computer. Dynamical addresses are also not taken twice into consideration.
- The effectiveness of rich content designed with Flash or Ajax can be measured.
- Once the tag is implemented, it is automatically executed by the browser without any assistance.
- The tags can be modified anytime to adjoin supplementary information about the user’s behavior (screen size, operating system, purchase, use of scrolling, access to special fields…).
- Improvable web pages of a domain can be inspected instead of the whole website, to get targeted information.

Compared to this, the log files have certain limitations:

- Without any application on the server, records and statistics cannot be produced.
- Before getting a web analytic report, you have to wait for the program to process the data of the server. It can sometimes take up to a few hours.
- Log files are not user-friendly, so one has to hire a network specialist to proceed the information.
• Log files size has the tendency to grow with the traffic, so they can take lots of place on the server.

• Log files can only identify IP-addresses and not unique visitors. In addition they cannot differentiate robot, spiders, from human users.

• Files can only be examined if they are on the same server as the one on which the analytical program is installed.

Packet analysis tools have the same downsides as those cited above. What is more, an additional device can complicate the maintenance and the architecture of the network. It can also be overloaded due the mass of information transmitted from the server farm and induce delays on the network. Moreover, as cookies, there are some privacy issues taking place, as information like passwords and identifiers can be intercepted. On the other hand, they can be associated with JavaScript tags to capture the potency of rich Internet applications, however solely a few web analytics companies implement this method.

The best solution is yet the mixture of genres with hybrid methods. Thus a company can use page tagging and still have a server-side program parsing its log files. Another method advises, instead of using a client-based tag, to implement a server-side tag that will be executed before the web page is sent to the client.

Now that we are aware of the benefit and disadvantages of these solutions, we have to determine which one to use. To this end, we tested a few web analytical tools.
11.1  Overview of the case study

This case study is targeted at evaluating the performance, accuracy, and capabilities of 9 different analytics packages which are implemented on our test website\(^6\). The goals of the project are as follows:

- Evaluate ease of implementation,
- Evaluate ease of use,
- Understand the basic capabilities of each package,
- Solve specific problems on each web site,
- Discover the unique strengths of each package,
- Discover the unique weaknesses of each package,
- Learn about the structural technology elements of each package that affect its capabilities,
- Learn how to better match a customer's needs to the right analytics package.

The analytic packages fall under two categories:

- Self-hosted analytic solutions;
- Third-party hosted analytic solutions.

\(^6\) http://shop.muazcisse.org/
11.1.1 Self-hosted analytic solutions

These are solutions are hosted on the server of the website being evaluated. The data collected is stored on the server. There is no third party involvement. Packages used in this study that fall under this strategy are:

BBCClone

BBclone is a web counter written in PHP and gives a detailed view of the visitors of your web site by displaying the nth last users (and the data they provided, like their IP, browser and so on) that visited the web site, not just showing a number of visitors. [BBclone, 2009] It enables webmasters and site owners to see who is visiting their sites, what browser people use, where they came from etc.

Figure 7: BBCClone admin interface

Setup

To install BBCClone on your website:
• Download the files from their website\(^7\);  
• Unzip the archive and upload the files to your web server;  
• Open the config.php file\(^8\) and make the necessary adjustments;  
• Simply add the activation code to the end of the php files you want to be counted

```php
define("BBC_PAGE_NAME", "Test");
define("_BBCLONE_DIR", "bbclone/");
define("COUNTER", _BBCLONE_DIR."mark_page.php");
if (is_readable(COUNTER)) include_once(COUNTER);
```

**Code 5 : Activation code source for BBClone**

After successfully setting up BBClone, it will immediately start to generate reports on your visits.

**ClickHeat**

ClickHeat is a visual heatmap of clicks on a HTML page, showing hot and cold click zones. [ClickHeat, 2009]

![ClickHeat admin interface](image)

**Figure 8: ClickHeat admin interface**

---

\(^7\) [http://bbclone.de/download.php](http://bbclone.de/download.php)  
\(^8\) This file is located in the folder bbclone/conf/
It enables site owners to see which section of a website is clicked.

Setup

To install ClickHeat on your website:

- Download the files from their website;  
- Unzip the archive and upload the files to your web server;  
- call the directory from your browser;  
- follow on-screen instructions  
- After the installation, simply add the activation code to the end of the files you want to be counted

```html
clickHeatSite = '';clickHeatGroup = 'index';clickHeatServer = your-server.com/clickheat/click.php';initClickHeat(); //-->
</script>
```

**Code 6** : Activation code source for ClickHeat

After successfully setting up ClickHeat, it will immediately start to generate reports on your visits.

**FireStats**

FireStats is a web statistics system which allows you to collect Real time statistics, recent referrers, search engine keywords, recent popular pages, etc. [FireStats, 2009]

---

9 http://www.labsmedia.com/clickheat/index.html

10 To visit the page, paste this address in your browser: http://www.yoursite.com/clickheat/index.php
Setup

To install FireStats on your website:

- Download the files from their website\(^{11}\);
- Unzip the archive and upload the files to your web server;
- create the tables inside your MySQL database;
- call the directory from your browser\(^{12}\);
- follow on-screen instructions

\(^{11}\) http://firestats.cc/wiki/Download

\(^{12}\) say http://your-server.com/firestats
After the installation, simply add the activation code to the end of the files you want to be counted.

Code 7: Activation code source for FireStats

After successfully setting up FireStats, it will immediately start to generate reports on your visits.

Grape Web Statistics

Grape is a free, open source program that allows web developers to keep accurate statistics of visitors. [Grape, 2009] It enables site owners to record unique and total website visits.

Setup

To install Grape on your website:

- Download the files from their website¹³;
- Unzip the archive and upload the files to your web server;
- Create the tables inside your MySQL database;
- Edit the values in quotation marks in the `/includes/connect.php` file;

---

¹³ http://www.quate.net/grape
• call the setup file from your browser\textsuperscript{14};

• follow on-screen instructions

• After the installation, simply add the activation code to the end of the files you want to be counted

\textless \texttt{script src="insert-\textit{location-to-grape}/?js" type="text/javascript">}  
\texttt{</script>}

\textbf{Code 8 : Activation code source for Grape}

After successfully setting up Grape, it will immediately start to generate reports on your visits.

\textbf{phpMyVisites}

phpMyVisites is a free and open source web statistics software. It enables website owners to see the keywords their visitors typed in the search engines (google, yahoo, live, etc.), know from which country they come from, which pages they are interested in, do they come often and how often, ... and much more information. [phpMyVisites, 2009]

\textbf{Figure 11: phpMyVisites admin interface}

\footnotesize{\textsuperscript{14} \url{http://www.your_web_site.com/includes/setup.php}}
Setup

To install phpMyVisites on your website:

- Download the files from their website\(^{15}\);
- Unzip the archive and upload the files to your web server;
- create the tables inside your MySQL database;
- call the directory from your browser\(^{16}\);
- follow on-screen instructions
- After the installation, simply add the activation code to the end of the files you want to be counted

\[
\text{Code 9 : Activation code source for phpMyVisites}
\]

After successfully setting up phpMyVisites, it will start to generate reports on your visits after 3 minutes [phpMyVisites, 2009].

\(^{15}\) http://www.phpmyvisites.us/downloads.html

\(^{16}\) http://www.your_web_site.com/phpmv2/
Piwik

Piwik is a downloadable, open source (GPL licensed) web analytics software program. It provides you with detailed reports on your website visitors: the search engines and keywords they used, the language they speak, your popular pages... and so much more. [Piwik, 2009]

![Piwik admin interface](https://example.com/piwik.png)

**Figure 12:** *Piwik admin interface*

**Setup**

To install Piwik on your website:

- Download the files from their website\(^\text{17}\);
- Unzip the archive and upload the files to your web server;
- create the tables inside your MySQL database;
- call the directory from your browser\(^\text{18}\);
- follow on-screen instructions

---

\(^{17}\) [http://piwik.org/](http://piwik.org/)

\(^{18}\) [http://yourdomain.com/piwik/](http://yourdomain.com/piwik/)
After the installation, simply add the activation code to the end of the files you want to be counted

```html
<!-- Piwik -->
<a href="http://piwik.org" title="Open source web analytics"
onclick="window.open(this.href);return(false);">
<script language="javascript" src="/yourdomain.com/piwik/piwik.js"
type="text/javascript"></script>
<script type="text/javascript">
piwik_action_name = '';
piwik_idsite = 1;
piwik_url = '/yourdomain.com/piwik/piwik.php';
piwik_log(piwik_action_name, piwik_idsite, piwik_url);
//-->
</script>
</a>
<!-- /Piwik -->
```

Code 10: Activation code source for Piwik

After successfully setting up Piwik, it will immediately start to generate reports on your visits.

**SlimStat**

SlimStat is a simple web stats analyser based on ShortStat\(^\text{19}\). It enables website owners to:

- Drill-down and filter data in any way you choose
- Shows visits and unique IPs, not just hits
- Option to hide visits from search engine crawlers
- Allows you to ignore specified IP address ranges
- Attempts to filter out spam [SlimStat, 2009]

\(^{19}\) [http://shortstat.shauninman.com/pact/](http://shortstat.shauninman.com/pact/)
Setup

To install SlimStat on your website:

- Download the files from their website\(^20\);
- Unzip the archive and upload the files to your web server;
- create the tables inside your MySQL database;
- call the directory from your browser\(^21\);

---

\(^20\) http://wettone.com/code/slimstat#download
follow on-screen instructions

After the installation, simply include the inc.stats.php file in your PHP wherever you would like stats to be counted. Use code similar to:

```
@include_once( $_SERVER["DOCUMENT_ROOT"]."/slimstat/inc.stats.php" );
```

**Code 11 : Activation code source for SlimStat**

After successfully setting up SlimStat, it will immediately start to generate reports on your visits.

After introducing the server-side methods, we will present the client-side applications.

### 11.1.2 Third-party hosted analytic solutions

**Clicky**

Clicky Web Analytics is a tool used to way to monitor, analyze, and react to your blog or web site’s traffic in real time. It enables website owners to Track downloads, Ajax and Flash events, See every action by every visitor among other things. [Clicky, 2009]

![Clicky admin interface](http://yourdomain.com/slimStat/)

**Figure 14: Clicky admin interface**
Setup

To install Clicky on your website:

- Register your account on their website\(^{22}\);
- After opening your account, simply include this script in your PHP wherever you would like stats to be counted. Use code similar to:

```
<a title="Clicky Web Analytics" href="http://getclicky.com/ Site ID ">
  <img alt="Clicky Web Analytics" src="http://static.getclicky.com/media/links/badge.gif" />
</a>
<script src="http://static.getclicky.com/ Site ID .js" type="text/javascript"></script>
<noscript>
  <img alt="Clicky" width="1" height="1" src="http://static.getclicky.com/ Site ID "/>
</noscript>
```

**Code 12**: Activation code source for Clicky

After setting up Clicky, it will immediately start to generate reports on your visits.

Google Analytics

Google Analytics allows website owners to gain rich insights into their website traffic with Advanced Segmentation, Custom Reporting, Motion Charts, and more. [Google Analytics, 2009]

This solution is used by the University of Fribourg to collect data on the number of visitors and the time they spend on the website.

---

\(^{22}\) [http://getclicky.com/user/register](http://getclicky.com/user/register)
Setup

To setup Google Analytics on your website:

- Register your account on their website;23
- After opening your account, simply include this script in your PHP wherever you would like stats to be counted. Use code similar to:

```javascript
<script type="text/javascript"> try { var pageTracker = _gat._getTracker("Site ID"); pageTracker._trackPageview(); } catch(err) {} </script>
```

**Code 13**: Activation code source for Google Analytics

After successfully setting up Google Analytics, it will immediately start to generate reports on your visits.

---

23 http://www.google.com/analytics/sign_up.html
After showing the different solutions we implemented on our website, we will compare their performance.

### 11.2 Key Findings

This section presents the final findings from using these analytical tools.

- Mere figures often deliver little information due to the differences in the packages. Therefore, specialists are needed to analyze and translate the figures into actual data so as to reveal the hidden information contained in these figures.

- Google Analytics offers the most complete set of functionalities.

- BBClone, Firestats and Grape do not give information on the average time visitors spend on a website.

- BBClone and Clicky do not give information the number of page views.

- Firestats and Slimstat do not give information on the number of Unique visitors.

---

<table>
<thead>
<tr>
<th>TOOL</th>
<th>VISITS</th>
<th>UNIQUE VISITORS</th>
<th>PAGE VIEWS</th>
<th>AVG TIME ON SITE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBClone</td>
<td>190</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FireStats</td>
<td>105</td>
<td></td>
<td></td>
<td>76</td>
</tr>
<tr>
<td>Grape</td>
<td>82</td>
<td>14</td>
<td></td>
<td>82</td>
</tr>
<tr>
<td>PhpMyVisites</td>
<td>19</td>
<td>10</td>
<td>80</td>
<td>3min 45sec</td>
</tr>
<tr>
<td>Piwik</td>
<td>52</td>
<td>40</td>
<td>261</td>
<td>4 min 7sec</td>
</tr>
<tr>
<td>SlimStat</td>
<td>66</td>
<td></td>
<td>804</td>
<td></td>
</tr>
<tr>
<td>Clicky</td>
<td>78</td>
<td>18</td>
<td></td>
<td>5min 54sec</td>
</tr>
<tr>
<td>Google</td>
<td>52</td>
<td>27</td>
<td>221</td>
<td>4min 28sec</td>
</tr>
</tbody>
</table>

**Tableau 2: Web analytics solutions benchmark**
- Web analytics packages, installed on the same web site and configured the same way, produce different numbers in their reports.

- By far the biggest source of error in analytics is implementation error. A Web analytics implementation needs to be tested to make sure it has been done correctly.

- Two other major factors drive differences in the results. One of these is the placement of JavaScript on the site, as being placed far down on a page may result in some users leaving the page before the JavaScript can execute. Traffic that is not counted as a result of the JavaScript can be considered an error, because the data for that visit is lost (or at least the data regarding the original landing page and, if the visitor came from the search engine, the keyword data would also be lost).

The other factor is differences in the definition of what each package is counting. The way that analytics packages count visitors and unique visitors is based on the concept of sessions. There are many design decisions made within an analytics package that will cause it to count sessions differently, and this has a profound impact on the reported numbers.

Note that this should not be considered a source of error. It is just that the packages are counting different things, equally well for the most part.

- To help address these accuracy problems, you should calibrate with other tools and measurement techniques when you can. This helps quantify the nature of any inaccuracies, and makes your analytics strategy more effective.
Conclusion

Data collection and analysis is the key factor to a successful customer relationship management. The first process aims at enabling market research and decision support. Furthermore it helps gathering data that conveys a certain idea of the company. With this information a later phase of review, in which the valid information will be interpreted, takes places. Thus, appropriate changes can be made according to a continuous improvement cycle, so that the web site will be able to attract more visitors, we seek to transform into consumers.

The objectives being set, it is then to implement both processes with informatics tools. The communication between the client and its server only lasting for the time of a session, information to identify visitors are lost. Other means had to be found to fill this gap. Among these, we notably find the server-side and the network analysis methods that employ the web logs to trace the route of visitors. On the other hand, web designers can use client-side tools like page tagging or web beacon, based on cookies.

Both techniques have advantages and drawbacks. Usually server-sides methods are more reliable, since the same device records the requests and the traffic. They are already freely available and installed on the network, the information is examined in the enterprise and not by a third party. Furthermore log files register successful and failed downloads and requests. Nevertheless these files are not user friendly, so one has to resort to web analysts to examine them. Besides they need to be processed depends on their size, so it can take lots of time when the site is successful.

Client-side solutions do not have these disadvantages. One does not even need a server to use them. They are obtained faster, due to the fact that the information is automatically processed instead of being recorded. In addition the cookies are parameterizable: they can capture specific information about the traffic and differentiate unique visitors using the same IP address. Nonetheless, a monthly fee has to be paid to the Application Service Provider, even if data can be lost due to storage space or network problems. If the company wants to change its analysis program, switching costs have to be taken into consideration. What is more if the visitor’s browser does not enable JavaScript information can be lost.

Even if using web analytics can seem at first not so difficult to implement, it takes a lot of organization. Analysts need to define the goals to be achieved and select the most important metrics. The work force’s actions have also to be coordinate, i.e. employees from the marketing, IT, administration department will have to work together, accepting on the contrary of their respective field, that the information gathered cannot give an exact representation of the reality.
At the end of the day, only avenues of research but no clear responses are provided. One has to avoid the temptation to change too quickly from a solution to another in case of deception, or modify the website too quickly without leaving the customer time to respond to new inputs. A company has to work under these constraints to best fulfill the needs of their customers, to hold its own game.
Bibliography

[Apache, 2009]

[BBClone, 2009]

[Clicky Web Analytics, 2009]

[ClickHeat, 2009]

[Clifton, 2008]

[Dosh Dosh, 2009]

[FireStats, 2009]

[Google Analytics, 2009]

[Grape Web Statistics, 2009]

[Hassler 2008]

[Kaushik, 2007]
[Kaushik and Waisberg, 2009]

[LeClaire and Sostre, 2007]

[Linklove, 2009]

[Massot, 1999]

[phpMyVisites, 2009]

[Piwik, 2009]

[SlimStat, 2009]

[WebbyAwards, 2009]