A Novel Appraoch for Comparing Third Party Web Analytical Tools for General Data Protection Regulation Policy

Veena Dhari, PhD
Department of Computer Science & Engineering
Rabindranath Tagore University
Bhopal
Madhya Pradesh - 462026

Meenakshi Garg
Department of MCA
Vivekanand Education Society Institute Of Technology Mumbai
Maharashtra 400071

ABSTRACT
This study seeks to explore the potentials and limitations of using various Analytical approaches and tools like google analytics, statcounter and web log files to analyze the behavior of online users in the context of content based websites and how Eurpean Union user consent policy has affected this data collection. Using these analytical tools this study will focus on type and amount of data gathered by these tools. The findings are expected to demonstrate useful features of Google Analytics, log files and other visualization tools and extend our understanding of online users. This study will also focus on the design and development of online educational resources and websites for collegiate contexts.

General Terms
Data analysis, pattern recognition, data privacy, Learning Analytics, Online Learning, data collection, personally identifiable information (PII).

Keywords
EU consent policy, web analytics, Google analytics, Stat counter, web log files, General Data Protection Regulation (GDPR)

1. INTRODUCTION
Today billions of users visit voluminous websites for consumer data , monetary management, educational information and for lots of other different services.[1] Once a client visits the websites, he leaves massive data within the log files and his activity can also be tracked by many analytical tools. Storing information in log files becomes helpful only if it's analyzed in meaningful manner for future business growth. Web mining is done to done to extract usage patterns from the user's behavior. It has become necessary for webmasters to utilize data mining techniques to mine out the required data resources, and to trace and analyze their usage patterns. These factors create the need of designing server side and client side intelligent systems which will effectively mine information especially for the content based websites.

Web analytics captures individuals' data who seek usage of information on the Internet [2], and it can also uncover key data about site clients and their online practices, for example, the essential socio economics areas, demographics, interests, navigation patterns etc. Subsequently, web analytics has picked up an expanding measure of consideration from associations, for example, business companies and government organizations educational institutions with an objective to meet organization goals. [3] Web analytics has additionally been applied in the field of online education, and is regularly alluded to as learning process when its information center is solely around the students, learning procedure, and learning contexts. Learning analytics powerfully catches and drills down user created information from sources, for example, the Internet, PCs, cell phones, and learning the board frameworks (LMS), and in this way "give significant knowledge into what is really occurring in the learning procedure and recommend manners by which webmasters can make enhancements".[4]

2. GENERAL DATA PROTECTION REGULATION
General Data Protection Regulation, or GDPR, the new EU data protection regulation was implemented on May 25 2018 and it was designed to modernize laws that protect the personal information of individuals. Cookies were also included in the scope of online identifiers. GDPR states that all cookies – even pseudonymous ones – can be considered personal data if there is any potential to use them to single out or identify an individual. [5

The General Data Protection Regulation ("GDPR") is a legal framework system that expects organizations to secure the individual data and security of European Union (EU) natives for exchanges that happen inside EU part states. It covers all organizations that deal with the data of EU residents, explicitly banks, websites, insurance agencies, and other money related organizations.

The full content of GDPR is contained 99 articles, setting out the privileges of people and put commitments on organizations that are liable to the regulation. GDPR's arrangements likewise necessitate that if any data sent outside the EU is secured and directed. At the end of the day, if any European citizen's data is collected, organization should be agreeable with the GDPR.

It is an elevated requirement to meet, necessitating that organizations contribute substantial wholes of cash to guarantee they comply with GDPR. As indicated by the EU's GDPR site, the enactment is intended to "blend" data security laws crosswise over Europe, giving more prominent protection and rights to people.

With the enforcement of GDPR, two noteworthy defensive rights ought to be featured. Initially, the privilege of deletion or the privilege to be overlooked should be taken care. If a European citizen doesn't want his information to be stored out there he has the privilege to ask for its evacuation or deletion. Second, the privilege of convenience should also be considered. With regards to "opt in/opt out " provisions, the notification to clients must be clear and exact as to its terms.[6]

GDPR requires clear consent and justification. Compliant with the GDPR, the accompanying kinds of information is tended to and secured: [7]

i. Personally identifiable information, including names, addresses, date of births, social security numbers

ii. Web-based data, including user location, IP address, cookies, and RFID tags

iii. Health (HIPAA) and genetic data

iv. Biometric data

v. Racial and/or ethnic data

vi. Political opinions

vii. Sexual orientation

2.1 Web Analytics Risk Assessment

Regardless of the European inception and focal point of these standards, the enactment contains genuine ramifications (20 million Euros, or 4% yearly income if that is more noteworthy!) for any association which gathers, procedures, or stores client information. The extent of this enactment did not depend on EU citizenship. Or maybe, it covers any client getting to a website or application from any area inside the EU limits. [7]

Level of risk for collecting data through Web Analytics in European depends on the:

i. The size of your company

ii. Business Owner location

iii. User base

iv. Users locations

Table 1. GDPR Web Analytics Risk Assessment

Risk	High	Moderate	Low
Business Owner located in any European Union Country	✓		
Is website is visited by EU residents	✓		
Is company achieving high volume sales?	✓		
If any personally identifiable information is collected in web analytics?	✓		
Is user IDs being tracked		✓	
Is data being collected for remarketing purposes?		✓	
None of the above			✓

3. WEB ANALYTICS TOOLS

Web Analytics tools are intended to make website analysis simple. The idea is to assemble, compose precisely, and present information regarding a webpage's performance.

In this study we are going to present analysis of few analytical tools like Google Analytics, Stat counter and web log files wrt to General Data Protection Regulation (GDPR).

3.1 Stat Counter

Stat Counter is a straight forward web analytical tool that uses simple JavaScript to track, break down and comprehend website visitors. [8] [9]

Fig 1: Stat Counter Overview

3.2 Google Analytics

Google Analytics is a free Web service that provides statistics and analytical insights for marketing purpose. This framework gathers information through a JavaScript page tag embedded in the code of pages. The page label works as a Web bug to assemble visitor's data. Nonetheless, in light of the fact that it's dependent on cookies to collect data, the framework can't gather information for clients who have crippled them[10]. Google additionally utilizes sampling in its reports instead of actual data.[11]

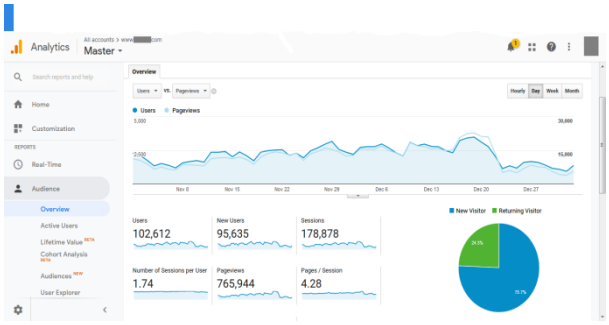

Fig 2: Google Analytics Overview

3.3 Web Log Files

Web log files are created by server automatically. Each "hit" to the Web webpage, including each perspective of a HTML report, picture or other article, is logged. The crude web log record is basically one line of content for each hit to the site. This contains data about who was visiting the webpage, where they originated from, and precisely what they were doing on the site. [12]

Here is a sample of log entry in Apache Combined format:

233.132.130.9 - - [15/May/2018:19:21:49 -0400] "GET /XYZ.htm HTTP/1.1" 200 9955 "https://www.XYZ.com/download.htm" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; Q312461)"

- 233.132.130.9- IP address
- "-"(hyphen) indicates Anonymous user id
- 5/May/2018:19:21:49- Web page access time
- -0400- The time zone
- GET/HTTP- HTTP request method
- 200- HTTP status code
- 9955- Number of bytes transmitted

4. SUMMARY OF TOOLS

This study compares the major aspects of three approaches via Google Analytics, Stat Counter and weblogs to check if they are General Data Protection Regulation compliant. In Table2 Y means that it is GDPR complaint, N means it's not complaint and NA shows that particular information is not maintained by that tool.

Table 2: Comparison of Various Analytical tools

Tracking Information	GDPR Complaint		
	1	**2**	**3**
Device tracking	Y	Y	Y
IP address tracking	N	NA	N
Visitors Location	N	N	NA
Referring URL	Y	Y	NA
Operating System and Browser	Y	Y	Y
Internet Service Provider	Y	Y	NA
Screen Resolution	Y	NA	NA
Visitor Path	Y	Y	Y
Time of visiting page	Y	NA	Y
Keywords used to search page in search engines	Y*	Y*	NA
Referrer source for a webpage	Y	Y	NA
Download Links	Y	NA	Y
Entry , Exit Pages & Exit Links	Y	Y	Y
Number of Page Views, Unique Visits and Returning Visits	Y	Y	Y
Site Speed	NA	Y	NA
Traffic from Social Media	NA	Y	NA
Requested Page size	NA	NA	Y

* Selective data is shown

1- Stat Counter, 2- Google Analytics, 3- Web Logs

5. CONCLUSION

The Google Analytics has been shown to be a useful tool for analytics and prediction that complies with GDPR. There are huge practical benefits of a Google Analytics as it helps in monitoring data in real time as well as provides deep insight into historical data. It also helps in quick visualization of data that can be helpful for effective business decision making. It contrasts with Stat counter and web log files as they store certain information that is not GDPR compliant and also real time monitoring is not possible.

6. REFERENCES

[1] Viktor Artemenko 2013 WebAnalytics in e-Learning : Agent Based and Neural Network Approach Lviv Academy of Commerce, 10 Tugan-Baranovskogo street, 79005, Ukraine

[2] Himani Singal , Shruti Kohli, 2014 Conceptual model for obfuscated TRUST induced from web analytics data for content-driven websites. Birla Institute of Technology , Mesra

[3] Jonathan R. Mayer, John C. Mitchell 2012 Third-Party Web Tracking: Policy and Technology. Stanford University Stanford, CA

[4] Yuh Jong Hu 2014 Privacy-Preserving WebID Analytics on the Decentralized Policy-Aware Social Web

[5] Raphaël Gellert 2018 Understanding the notion of risk in the General Data Protection Regulation Tilburg University

[6] Wanda Presthusa, Hanne Sørumb 2018 Are Consumers Concerned About Privacy? An Online Survey Emphasizing the General Data Protection Regulation Tavel. Westerdals Department of Technology, Kristiania University College, Christian Krohgs gate 32, 0186 Oslo, Norway

[7] General Data Protection Regulation https://gdpr-info.eu/ Access date 15 December 2018

[8] TerryBallard 2014 iGoogle and other useful products https://doi.org/10.1016/B978-1-84334-677-7.50005-X

[9] Web Analytics made easy https://statcounter.com/- Access date- 15 December 2018

[10] Sanda-Maria Dragos¸ 2011 Why google analytics can not be used for educational web content "Babes-Bolyai" University

[11] Google Analytics https://analytics.google.com/analytics/web/provision/#/provision Access date- 20 December 2018

[12] Chaitra L Mugali, AyeshaAzeema Maniyar, Padma Dandannavar 2015 Pre-Processing and Analysis of Web Server Logs KLS Gogte Institute of Technology