**Project title: Web site traffic analysis**

**Phase 1: Problem Definition and Design Thinking**

In this part you will need to understand the problem statement and create a document on what have you understood and how will you proceed ahead with solving the problem. Please think on a design and present in form of a document.

**Project Definition:**The project involves analyzing website traffic data to gain insights into user behavior, popular pages, and traffic sources. The goal is to help website owners enhance the user experience by understanding how visitors interact with the site. This project encompasses defining the analysis objectives, collecting website traffic data, using IBM Cognos for data visualization, and integrating Python code for advanced analysis.

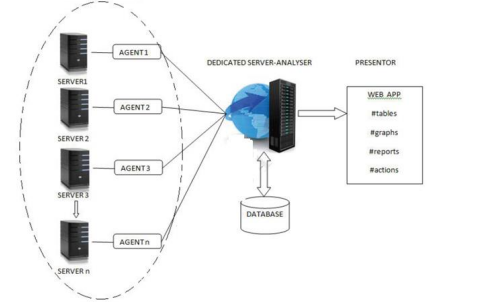
**Design Thinking:**

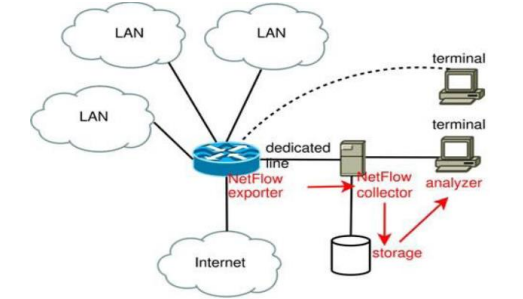
1. **Analysis Objectives:**

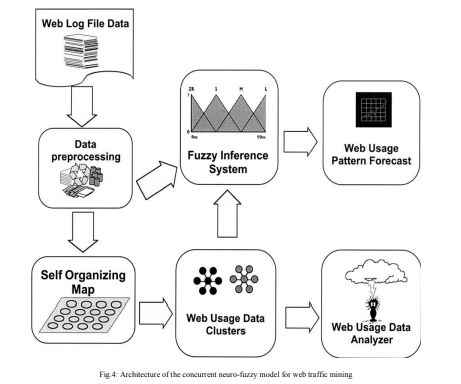
With the rapid increasing popularity of the WWW, Websites are playing a crucial role to convey knowledge and information to the end users. Discovering hidden and meaningful information about web user’s usage patterns is critical to determine effective marketing strategies to optimize the Web server usage for accommodating future growth. Most of the currently available Web server traffic analysis tools explicitly provide statistical information. The web server traffic analysis tools make the use of Web Access Logs that are generated on the server while the user is accessing the website. A Web access log comprises of various entries like the name of the user, his IP address, number of bytes transferred timestamp etc. The task of web traffic analysis tools becomes more challenging when the web traffic volume is enormous and keeps on growing. In this paper, we propose a various model to discover and analyze useful knowledge from the available Web log data and also provides a comparative study of variety of Log Analyzer tools exist which helps in analyzing the traffic on web server.

1. **Data Collection**: The web traffic starts with the high-level activities such as clicking a link and increases with low-level activities such as travelling through network switches and cables. In other words, Web traffic is usually initiated by users via the use of web browsers. It begins with a click to access a URL. Traffic flow starts with a mouse click, which sends browser information to a server that uses predetermined rules and methods to obtain user browser requests. Based on these rules, the server then decides what action is needed. Now a days, the web traffic is increases enormously because there is continuous increase of Internet users each year is motivating online shop, gambling site, and botnet owners to take control of users‘ moves to point them to their sites. Therefore, there is need of web traffic analysis tools. These tools handle and categorize the traffic and increase the workload handling capacity of the web server. B. LOGS Web server logs stores click stream data which can be useful for web traffic analysis purposes [3]. They are plain text (ASCII) files which contain information about User Name, IP Address, Time Stamp, Access Request, URL that Referred, error codes (if any) etc. and generally reside in the web servers. Traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log [4]. The Transfer and the Agent Log are said to be standard whereas the error and referrer log are considered optional as they may not be turned on. Every log entry records the traversal from one page to another, storing user IP number and all the related information [5]. If logs are utilized properly, it can be very useful in turning the websites visitors into customers especially in case of an e-commerce website. It guides the analyst in determining the navigational pattern of the user i.e. which pages are frequently visited by the user, the kind of errors that user gets, etc. A variety of tools are available that take the web access logs as an input and generate the reports as an output. These tools provide us with all sorts of information starting from how many hits the site getting to the number of visitors accessing the site, the browsers that they use the length of their stay, and much more. Some of the tools that are available are: 1) Google Analytics: It is a free utility provided by Google which helps in keeping a track of unique visitors. It also helps in determining which marketing packages are offering the best .For using this tool, installation is not required, only requires a Google account. Email report facility is available in Google analytics. 2) AWStats: It is available free of cost. This tool works as a CGI Script or from command line. It displays all sorts of information that the log contains. 3) WebLog Expert: Yet another log analyzer tool that provides thorough analysis of the web access logs. It provides the users with specific and precise information about user‘s statistics. It supports log files extracted from Apache and IIS. For using this tool, there is no need of creating any account but profile creation is required. 4) Analog: This is an easy to use and install freely available log analysis tool. It is extremely fast, highly scalable, works on any operating system and an easy to install tool.

**Visualization:**







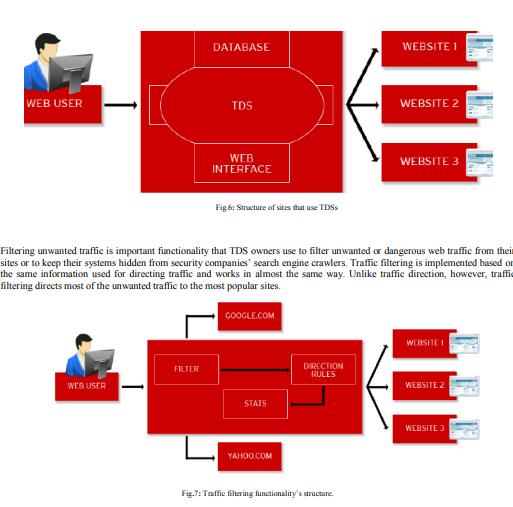


Fig.7 shows that web traffic is usually filtered then directed from the very beginning while writing information to a database or a text file at the same time for future statistical means. TDS owners that use the traffic filtering functionality usually redirect unwanted requests to the most popular legitimate sites in order to hide their systems‘ main function. For example, security companies that scan lists of possibly malicious URLs and IP addresses are filtered by malicious TDS owners to block the HTTP requests of antivirus crawlers. This instead directs security experts to non-malicious legitimate sites such as Yahoo! and Google.