Identification of Anomalous Clients' Request by Analyzing Server Log File using Apache Hadoop Framework and Tableau

V. Bavathuja, S. Raahini, M. Ramashini and S.T.C.I. Wimaladharma

Department of Computer Science and Technology, Uva Wellassa University, Badulla, Sri Lanka

Information systems provide information about its state and operation in the form of log records. These records are composed of log entries containing information related to a specific event, which can be related to security. Potential security breaches can be revealed by analyzing log files and looking for anomalies that occurred at a certain time during the device operation. Log files from proxy server of Uva Wellassa University of Sri Lanka will be analyzed using Hadoop Framework and Apache Pig in order to identify anomalous clients' Request. Anomalous clients' request identification refers to the problem of finding pattern in data that do not conform to expected behavior. These non-conforming patterns are often referred to as anomalies, outliers or exceptions in different application domains. Log files of a proxy server are created and maintained by the server itself and analyzing theses files will offer a valuable insight into server usage while they can be used in various applications, such as detecting intrusions on the web. The log files will be stored in Hadoop Distributed File System. Data preprocessing and analyzation will be done using Apache Pig: a platform for analyzing large data sets. The analyzed data will be reported through Tableau dashboard. According to the research study, the total number of records after cleaning is 817,426 and 856 unique IP addresses have accessed the proxy server from the period of Thursday, 26 April 2018 01:14:48.138 to the period of Friday, 27 April 2018 10:31:23.834. Several findings including the total visits and bandwidth were found and displayed using graph and charts. This information along with other findings can be applied to find solutions for many legitimate problems such as, user/customer behavior analysis, etc.

Keywords: Proxy server, Log files, Apache hadoop, Apache pig, Tableau.