A STUDY ON THE BIG DATA LOG ANALYSIS: GOALS, CHALLENGES, ISSUES, AND TOOLS

[1]MOSTAFA MOHAMED SHENDI, [2]HATEM MOHAMED ELKADI,
[3]MOHAMED HELMY KHAFAGY

Information System Department, Cairo University, Egypt Computer Science Department, Fayoum University, Egypt
E-mail: mostafa_shendi@yahoo.com

Abstract - Security, availability, and performance are becoming more frequently requested and sophisticated. Traditional solutions are unable to protect organization's assets or keep their services running smoothly. These solutions need to focus more on customer needs and satisfaction. Organizations needs to perform real-time analysis on a huge amount of data from various types to discover anomalous fragments within a reasonable response time. Business can widen the scale of processed data, accelerate the speed of threat detection, keep their services up and running by monitoring the servers' status, predict failure before it happens and increase customer satisfaction by providing efficient service in a timely manner. Processing the huge amount of the system's log files using relational database technology has been facing a bottleneck. To analyze such large data sets we need parallel processing system and reliable data storage mechanism. Big Data is the solution to overcome these issues. The main purpose of this paper is to highlight the characteristics of Big Data and present a review of log file analysis in Big Data environment as a first step towards getting the maximum benefits of big data in logs analytics.

Keywords - Big data, Hadoop, MapReduce, HDFS, log analysis, log files.

I. INTRODUCTION

Nowadays many businesses are going online. Each field has its own way of bringing their applications and businesses online. We are seeking a mechanism to handle huge volumes of log files generated in different structures and from different devices and applications. Many studies and papers - discussed previously from different perspectives - highlights how big data can be useful in analyzing the logs and user's behaviors on a web page, and detecting anomalous behaviors, errors and exceptions.

Log files:
Logs are computer-generated files that capture network and server's operation data containing the whole information regarding the user's activity. These activities are written in various log files such as a weblog, firewall log, network log, Router log, etc. these log files have millions of entries. Log analysis needs a long time to be investigated.
Analysis of log files has been very important in resolving many issues [1]. Servers, routers and other devices logs may be our best line of defense.

Big Data
Basically, Big Data is defined as data sets that could not be perceived, acquired, managed and processed within a reasonable time by traditional IT and software or hardware tools. Big data exceeds the processing capacity of conventional database systems [2]. The data is hugeand massive, it moves at a very high speed, and does not have to fit in the structures of existing database architectures. To gain value from this data, there must be an alternative way to process it. Big data analytics is the key to unlock the insights from all data types as it enables us to analyze all structured, semi-structured and unstructured data together. It is powerful because it enables the organization to combine, integrate and analyze all data at once regardless of their source, type, size or format in order to generate the insights needed to address a wide range of business challenges.
Hadoop-MapReduce framework provides parallel distributed processing and reliable data storage for large volumes of log files. Here Hadoop's characteristic of moving computation to the data rather than moving data to computation helps to improve response time [3].
Results from the reviewed papers show that the majority of researchers applied MapReduce as the main component of Hadoop for analyzing the log files and HDFS as the data storage. Previous researchers have also used some tools and algorithms with the Hadoop Framework for analysis purposes. The output of this paper will provide a comprehensive review of Big Data usages in analyzing different types of log files with different methods, tools, and techniques.
Hadoop takes log files analysis to the next level by speeding and improving security forensics and providing a low-cost platform to detect compliance violations [4].

II. CONSIDERATIONS AND CHALLENGES INBIG DATA LOGS ANALYSIS

In the Last few years, big data has been acquired in several domains. A web-based application such as internet text, social computing, recommender systems, online communities, markets prediction, social network analysis, and internet search documents encounter big data frequently. Also, millions of records are generated from the devices over a network which should be reviewed and handled in appropriate response time.

To handle the challenges, we need to manage various computational and information security complexities and methods to analyze Logs. The logs are generated in different formats with different structures and contains useless data that need to be cleaned.
When working with large logs, practitioners often face issues such as scarce storage, incapable analysis tools, inaccurate capture and replay of logs, and inadequate privacy. Researchers have devised some practical solutions, but important challenges remain:

Data Distribution
Performing computation on large volumes of log files has been done earlier but what makes Hadoop different is its simplified programming model and its efficient, automatic distribution of data and portability [5].

Isolation of Processes
Each individual record is processed by a task isolated from other tasks, limiting the communication overhead between the processes by Hadoop. This makes the whole framework more reliable. In MapReduce, Mapper tasks process records in isolation. Individual node failure can be worked around by restarting tasks on another machine.

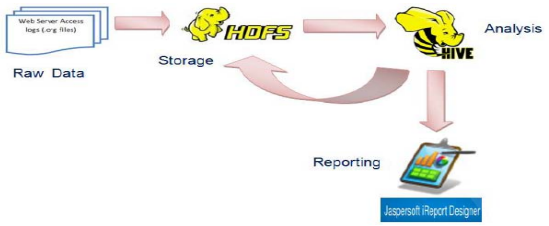
Type of Data
Log files are rows consisting of semi-structured or unstructured records. Hadoop is compatible with most types of data and more suitable even in case that the log files are mapped to a structured data due to RDBMS data size limitation. It works well for simple text files as well.

Fault Tolerance
Hadoop cluster solved the problem of
data loss. Blocks of the input file are replicated by a factor on multiple machines in the Hadoop cluster. Therefore, even if any machine goes down, another machine where the same block is residing will take care of further processing.

Data Locality and Network Bandwidth
Log files are spread across HDFS as blocks, therefore the node which operates on a subset of files will be detected by the locality of the node, and the purpose is reducing the strain on network bandwidth and avoid unnecessary network traffic [6].

III. LITERATURE REVIEW TECHNIQUES AND TOOLS

Numerous efforts have been expended to get the maximum benefits of using big data in the analysis of log files on different structures. Siddharth et Devesh [7] used a weblog analyzer tool for collecting the statistics over web sites, Pig scripts are written on the classified log files to satisfy the certain query. The log file is processed for exceptions and errors.
The tool worked on the basic information present in the log file as follows:
IP Address /Username - Time Stamp - Page Visited lastly - Success rate - User agent - URL - Request type
The paper used Map reduce technique to process the log files, Hadoop framework for parallel computation of log files, pig tool for taking a huge data from different sources and placing it into HDFS for further processing. The framework makes use of tableau tool for a pictorial representation of log files accessed by the users.
The limitation of the paper is the need to collect, analyze and gather events occurrences to subscribers such as operating system processes, database rules, and human operators.
Harish et Kavitha [8] applied Hadoop-MapReduce programming model for analyzing web server log files where data is stored on multiple nodes in a cluster so that access time required can be reduced and MapReduce work for large datasets giving efficient results to summarize the results for a web application, perform log analysis that improves the business strategies as well as generates statistical reports. The data cleansing is done using Hive query language. Only the logs with the status code are extracted and stored in Hive table for further analysis



Raw log data processing and visualizing
The paper analyzed website logs and obtained the following: General Statistics, Activity statistics, Access Statistics, Visits-per-Country and Errors. Then performed log analysis to improve the business strategies and to generate statistical reports. Which provides graphical reports showing hits for web pages, user's activity, and identifies in which part of the website users are interested, traffic sources, etc. Using these reports the business communities can evaluate parts of the website that needs to be improved, the potential customers and max hit traffic source.
Mayur Mahajan et Omkar Akolkar et. [9] Used web page collection algorithm that uses cluster mining to find a group of connected pages at a web site. The proposed algorithm takes web server access log as an input and maps it into form clustering. The output is obtained by mining the web user's logs. Faster analysis of real-time weblogs generated by various web sites with parallel processing using Apache Hadoop.
Recommendations will help the administrators improve the website in various locations according to

various factors. Using the analysis report the developers can fix the bugs in the website. Analyzing the real-time weblog which contains errors and exceptions and the number of times it has occurred.
The paper builds its system architecture which consists of Third Party Website, Web Log Analyzer, and User-interface
The system analyzes the log files and presents it to the user in a more understandable form such as pie-charts and bar-graphs, and gives recommendations to the user for improvement of the website for popularity and reputation risk matters.
The limitation of this work is that real-time weblog analysis needs to be done. The fluctuations can be shown in reports online as per the changes of the log.
Sayalee Narkhede et Tripti Baraskar [10] the structure used in the work consists of web servers on cloud Framework, Hadoop storage and MapReduce programming model and user interface.
Log files are collected from the web servers. HMR log processor is implemented in three phases: log pre-processing, interacting with HDFS and map-reduce programming model implementation.
HMR log file analysis tool provides graphical reports showing hits for web pages and it`s sub-elements, user's activity, and traffic sources. Data get stored in the hierarchy on several nodes in a cluster so that access time required can be reduced which saves much of the processing time. Here Hadoop's characteristic of moving computation to the data rather moving data to computation helps to improve response time. Secondly, MapReduce successfully works for large datasets giving efficient results.
Kyung et Jeong et Hyun [11] suggested the need for new Enterprise Security Management (ESM) which is used for integrated management of network systems such as firewall, IPS, VPN, etc. also they have mentioned that the current ESM has a limitation of blocking external cyber-attacks only due to using the networking attack detection method which monitors the traffic inflows from outside to inside.
The research uses big data to enhance the intelligence of security by analyzing the relationship between security and data events created from network, system, application service of main IT infrastructure. They collect and analyze rapidly a large amount of structured/unstructured data by distribution-based saving/processing technology through security log system using Big data. Fetch real-time analysis performance by distribution-based multi-searching, presenting a flexible dashboard for visible analysis, and applying correlation analysis technique of all elements. The integrated security control will be possible by securing availability and expandability in case of applying security log system using Big data.
Data collection Algorithm is expected to enhance customer satisfaction by restricting the inflow of harmful code developed in-house, and real-time monitoring becomes possible. The limitation of the research is the absence of any experimental setup and not presenting any sample results of the proposed work.
Sandesh et Abhishek et... [12] Discussed the vulnerability assessment and presented some free tools to use for smaller modules, list of some tools which are easy to use, so that anyone with basic technical knowledge can use them, allowing even a small enterprise to utilize them to generate results and take appropriate actions: Vulnerability, Threat, Attack and Exploit described different techniques to collect and analyze logs, which are a very important source of information and could be used for various security purposes such as a NetFlow, IDS, firewall log activity, file access, system errors, or login failures.
The limitation of the work was the fact that the researchers didn't propose a complete integrated module from the tools they have mentioned in order to be used by the SME`s companies.

Sharma et Arun [13] focused on the Web applications which uses a combination of server scripts (ASP, PHP, etc.) and client-side scripts (HTML, JavaScript, etc.) to develop the application. The script on the client-side communicates with the presentation of information, while the server-side script handles all the complex operation such as storing and retrieving the information. The paper added a mining layer of the web which is divided into three categories: web content mining, web structure mining and web usage mining to discover hidden information on the Web by data cleaning, filtration, data cube construction, analytical processing, and data mining.
The basic information found in the log file: User name, Path Traversed, Success rate, User Agent and URL.
The limitation of the paper is that there is no integration of the proposed work. Therefore, the paper only added a data mining step for data cleansing and filtration and didn't mention the enhancement percentage through it.

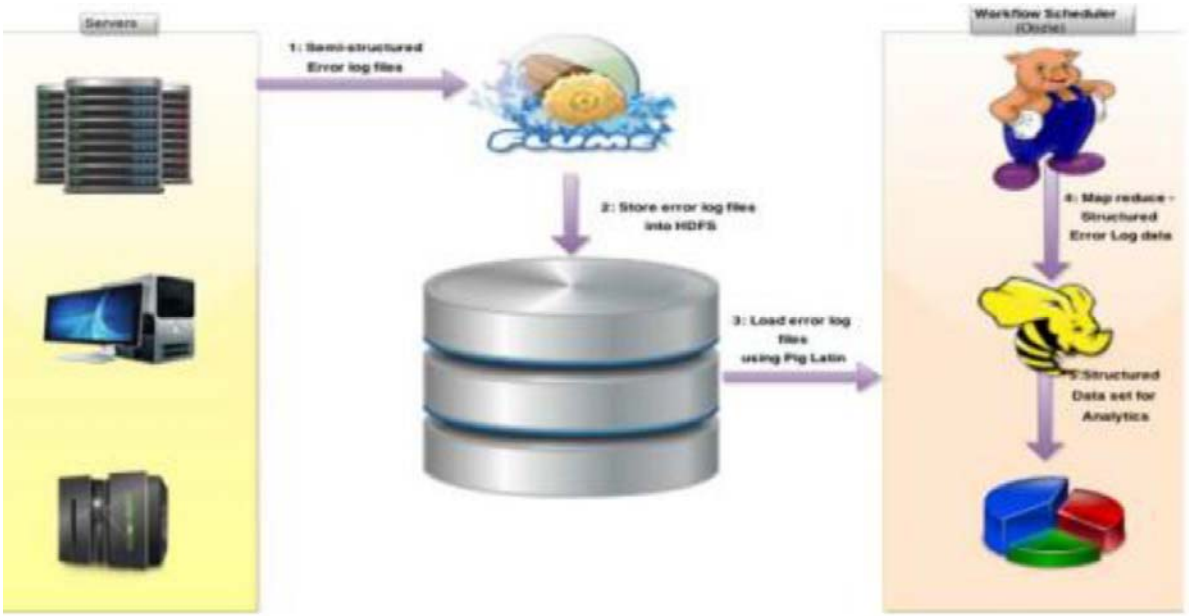
Rotsnarani et Mrutyunjaya [14] used Hadoop framework, data mining techniques to extract useful information and clustering method Centroid-based (K-Means Clustering) to cluster the input logs and then process them in Hadoop framework. The analysis result can be used for detecting security threats in the network and to inform the administrators about Intrusion Detection, SQL-Injection, system error, etc.
Log files can be classified into categories depending on the location of their storage. This classification can provide batch analysis in minimum response time and in-memory computing capacity to process logs in an efficient and stable way.
The paper focused on two types of logs: web server logs and application server logs.
The proposed technique consists of scanning logs in real time as well as in off time which can give the

best result of analysis of the logs that can be investigated by the IT admins to take preventive actions for the attacks.

The real-time logs can be stored and processed by using flume. In-memory engine for analytics is a better option for faster processing to detect attacks like SQL injection, Cross-Site Scripting (XSS), URL Injection, etc. The limitation of the paper is focusing on the web servers and applications logs only, it would be better to include network logs such as firewall and switches logs, etc... to the proposed work

Souza1 et Girish [15] Error Log Analytics discussed the conversion of data from semi-structured to a uniform structured format, such that Analytics can be performed over it. Business Intelligence (BI) functions such as Predictive Analytics are used to predict and forecast the future status of the application based on the current scenario. Proactive measures can be taken rather than reactive measures to ensure efficient maintainability of the applications and the devices.

Error Log is a file that is created during data processing to hold data known to contain errors and warnings. It is usually printed after completion of processing so that the errors can be rectified



After ETL operations performed, the error log file is brought into a uniform homogeneous format. To perform analytics on this data, it is loaded into a warehouse. Obtaining the sum of errors and warnings on the severity attribute is achieved using the Grouping operation. Cubing operation generates aggregates for all combinations of values in the selected columns. Historical error log data is maintained in the warehouse, which is used in various Analytics and business intelligence techniques.

Bhawna et Kiran [16] review paper explains how Big Data is analyzed by using the technique of Hadoop and why the Big Data Security Analytics is important to mitigate the security threats.

There should be several opportunities for big data security analytics to enter the enterprise security mainstream because of: continuing problems with detection and response of threats, Moore's law and open source and Tons of activity on the supply side.

Researchers must meet the needs of enterprises for developing a high-quality product by adopting security measures with the help of big data analytics with Hadoop. The paper uses big data analytics for analyzing enterprise data, and builds a framework based on Hadoop for dealing the targeted attacks using Big Data security analytics.

The limitation of the paper was the absence of any implementation for the proposed work, also they lacked focusing on the enterprise data security and developing security alerts which will provide employees with the ability to view the activity.

Jangla et Deepa [17] propose a new model based on big data analysis technology to prevent and detect previously unknown attacks. Explain attack predictions earned from applying big data technologies such as classification, text mining, clustering, and association rules.

Propose a system model that uses big data analysis technology for extracting data from various sources to react to previously unknown attacks.

Big data analysis uses various existing analysis techniques such as machine learning, artificial intelligence, data mining, etc.

The limitation of the paper was that it lacked the software ability to detect advanced threats, reacting accordingly and developing preventive measures for the future. The paper researchers should improve the quality and reliability of the security system.

Work / Year	Method	Result	Strengths	Weakness
Siddharth et Saraf. "Analysis of Log Dataand Statistics Report Generation Using Hadoop" [2015]	-Hadoop-MapReduce & Pig scripts - Pig scripts are written on the classified log files to satisfy a certain query. the log file is processed for exceptions and errors.	The framework makes use of tableau tool for a pictorial representation of log files accessed by the users.	finding the statics of web sites: IP Address /Username - Time Stamp - Page Visited lastly - Success rate - User agent - URL - Request type	- Just focus on the website logs. - Need to collect, analyze and signal events occurrences to subscribers such as operating system processes, active database rules as well as human operators.
Harish et Kavitha. " Statistical Analysis of Web Server Logs Using Apache Hive in Hadoop Framework " [2015]	- Hadoop-MapReduce & Hive - data cleansing is done using Hive query language	obtained: General Statistics, Activity statistics, Access Statistics, Visits-per-Country, and Errors	evaluate which parts in the website that needs to be improved, and whom are the potential customers, from which geographical region the website is getting maximum hits, etc., which will help in designing future marketing plans	- focus on application logs only - Doesn't have the capability of distinguishing between bad connections, called intrusions or attacks, and good (normal) ones
Mayur et Omkar. "Real-Time Web Log Analysis and Hadoop for Data Analytics on Large Web Logs" [2016]	-Hadoop-MapReduce &cluster mining -Webpage Collection algorithm that uses cluster mining to find a group of connected pages at a web site.	Provide analysis errors report which help the developer fix the bugs in the website	Help the administrator for the improvements of the website in various locations on various factors	- Real-time weblog analysis needs to be done. The fluctuations can be shown in reports online as per the changes of log
Sayalee et Tripti. "HMR LOG ANALYZER: ANALYZE WEB APPLICATION LOGS OVER HADOOP MapReduce" [2013]	-Hadoop-MapReduce &Interface - Builds its structure which Consists of major components: Web servers, Cloud Framework implementing Hadoop storage and MapReduce programming model and user interface.	provide graphical reports showing hits for web pages, user's activity, in which part of website users are interested, traffic sources, etc.	Business communities can evaluate which parts of the website need to be improved, which are the potential customers, from which geographical region website is getting maximum hits, etc.	- Limitation of the work scope, need to build the structure on real-time methods - Security filed should be included to classify the coming requests as normal or deny of service.
Kyung et Jeong "A Study on the Big Data Log Analysis for Security" [2016]	- Hadoop-MapReduce & NOSQL (Hive-Mongo database), Mahout-R language - collecting, saving, processing, and analyzing techniques based on intelligent information analysis platform for system construction of security log analysis using Big data	SIEM (Security Information & Event management)" intelligent security by integration with Big data analysis technique, such as the new generation security information analysis technique	Supporting fast searching and showing visualized method to analyze the logs. enhance customer service satisfaction by inflow harmful code in-house, and real-time monitoring becomes doable	- Limitation of selecting a part of the security field among diverse big data methodology. - The extra study of analysis techniques of big data analysis area can now be applied to diverse fields.

Sandesh et Abhishek "Survey on Log Analysis and Management" [2015]	- Hadoop-MapReduce & Splunk – Hive - Chukwa - discusses the vulnerability assessment utilizing some free and easy to use tools for smaller modules.	described different techniques to collect and analyze logs, which are a very important source of information and could be used for various security purposes such as NetFlow, IDS detection, firewall log activity, file access, system error, or login failure	Discusses the vulnerability assessment utilizing some free and simple tools for smaller modules so that even a small enterprise can utilize them to generate results and take appropriate actions.	They haven't implemented their solution but discussed the tools only, there is no future architecture or design
Sharma et Jhapate "Analytics of Web Log File through Map Reduce and Hadoop" [2016]	- Hadoop-MapReduce& Data Mining, Cloud -Use Web content mining, Web structure mining, web usage mining and, log analysis. analyzed data generation from different sites in massive amount is progressive and distributed over the cloud	Review the basics of log analysis as big data in the web environment. collection of web access information for web pages	elaborates the concept of Hadoop a Big data tool which is made for Handling big data analytics and meets with the expectation of ever growing demands of the data.	- There is no integration on this work has been known
Girija et Sanket "Big Data Analysis of Fetched Logs using Hadoop Framework" [2017]	- -Hadoop-MapReduce& Data mining (Centroid-based clustering (K-Means Clustering), Density-based clustering) -Hadoop framework to process huge amount of data extract useful information data mining technique is used, Various types of clustering like Connectivity-based clustering,	scanning logs in real time as well as in off time which can give the best result of analysis of logs that can be investigated by the IT admins to take preventive actions for the attacks	The collective result of the real-time and manually performed analysis will be then displayed on the web page for effective and easy interaction	- There is no implemented framework, only scanning of some related work
Souza et Girish "Error Log Analytics using Big Data and MapReduce" [2015]	- Hadoop-MapReduce& Pig, Flume, Hive, Oozie, R, Business Intelligence (BI), Data Mining, Machine Learning, Predictive Analytics Business Intelligence (BI) Predictive Analytics is used to predict and forecast the future status of the application based on the current scenario	Categories of severity are Error, Warning. And Info. -Abbreviations and Acronyms -Data Preprocessing Cleansing and Integration Pattern -Discovery and Analysis-Predictive Analytics -. Proactive measures to ensure efficient maintainability of the applications and the device	predict the future severity value Whenever message IDs are the values displayed in the left branch of the tree, then there is a high chance of the system going to an Error	- Doesn't perform direct analysis on the targeted logs - Create Error Log file during data processing on logs - No integration between targeted logs, each log is a separate processed
Bhawna et Kiran "Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data" [2015]	Hadoop-MapReduce & HBase, PIG, HIVE, SQOOP and ZOOKEEPER Review paper discuss how Big Data is analyzed by using the technique of Hadoop	design to meet the production needs of enterprises for developing a high-quality product by applying security measures with the help of Big Data Analytics with Hadoop	Plan to take up data collection, pretreatment, integration, map reduce and prediction using machine learning techniques.	- they have a challenge to implement and call for strong support in terms of security policy formulation and mechanisms
Jangla et Deepa "Development of an intrusion detection system based on Big data for detecting unknown attacks" [2016]	- Hadoop& combiner, Partitioned HBase, PIG, HIVE, SQOOP - Propose a new model based on big data analysis technology to prevent and detect previously unknown attacks	explain attack predictions earned from applying big data technologies such as classification, text mining, clustering, and association rules	system model that uses big data analysis technology for extracting data from various sources to react to previously unknown attacks	- Need to improve the ability of software to find advanced threats, react accordingly and develop preventive measures for the future.

Table-1 summarized the related research efforts of log analysis using Hadoop Framework. The proposed frameworks involving various types of log analysis proved that Hadoop can handle a variety of data. Mainly, the research work applied MapReduce as the main component of Hadoop for analyzing the log files and HDFS as the data storage. To fulfill the analysis purposes, the researchers have also used other tools and algorithms together with the Hadoop Framework. The results of these researches show that implementing Hadoop framework enables us to successfully minimize the analysis process response time as well as analyzing larger datasets.

IV. DISCUSSION (GOALS OF RESEARCH ON BIG DATA LOGS ANALYTICS)

The analysis of log files in real time and off time plays an important role in any IT infrastructure. One of the benefits is to make them robust enough to withstand against the hackers' attacks, it collects and analyzes logs, could also be used for various security purposes such as NetFlow, IDS, firewall log activity, file access, system error, or login failure and detecting security threats in network and alerting the administrators with the intrusion detection, SQL-Injection, system error, etc. One additional benefit is to enhance the intelligence of security by analyzing the relationship between security and data events created from network, system, application service of main IT infrastructure

The analysis reports help the developers fix the bugs in the website and handle exceptions and errors and capture the number of times they have occurred. Furthermore, it gives recommendations which will help the administrators to improve the environment in various locations on various factors.

The information about site visitors and their user activities, accessed files, paths in the sites, user navigation between pages, various browser statistics, and operating system statistics will help in studying user behaviors when using highly interactive systems and improve the business strategies as well as to generate statistical reports.

The visualization tool provides graphical reports showing hits for web pages, user's activity, traffic sources, in which part of the website users are interested, the potential customers and from which geographical region website is getting maximum hits, etc., which will help in designing future marketing plans. The previous table shows that most of the related researches uses Hadoop and MapReduce techniques. They focus mainly on logs analysis and extract useful information.

The output of the review is in the form of the graphical reports and charts to allow the administrator to adjust their websites.

Research techniques have strength in:

- Fetching the statistics of web sites
- Help administrators for the improvement of websites and handling the exceptions
- Provide graphical reports showing hits for web pages, user's activity
- But still lack important features such as:
- Real-time analytics can provide a better result and corrective actions.
- Scanning different logs at the same time which is very important, this type of integration may lead to more evidence.
- Analyzing log network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view
- Prediction model should be built, not just classify the events.

We recommend the research to target deficiencies such as handling the large volumes of data and different types of logs generated frequently, selecting a part of security field and provide the corrective actions immediately by using Hadoop, real-time analytics techniques and build a predictive model.

V. CONCLUSIONS

This paper presents a comparative study about the log data analytics from the traditional data analysis to the recent big data analysis. The main objective was to show how to get the maximum benefits of big data in logs analytics with different methods, tools, and techniques.

It was found that there is a limitation of selecting a part of the security field among diverse big data methodology, there is no integration between different security devices and a shortage in alerting methods.

Future work is to design a framework that includes ingesting and Processing Net Flow Data, alerting, flagging, transforms, filtering, Real-Time Processing - aggregations, windowed computations, and machine learning to be presented in different ways not only for detection of the failure and anomalous attack on time, but also to predict it.

We plan to implement such a predictive model which is capable of distinguishing between bad connections (called intrusions or attacks) and good (normal) ones, apply the module on different devices logs and integrate it to achieve maximum security, availability, performance and customer satisfaction possible.

REFERENCES

[1] Latib, M. A., Ismail, S. A., Sarkan, H. M., & Mohd Yusoff, R. C. (2015). Analyzing log in big data environment: A review. ARPN Journal of Engineering and Applied Sciences, 10(23), 17777–17784.

[2] Grable, J., & Lyons, A. (2018). An Introduction to Big Data. Journal of Financial Service Professionals, 72 No. 5(September), 17–20. https://doi.org/10.1016/B978-0-12-801967-2.00001-X

[3] S. Harish and G. Kavitha, "Statistical Analysis of Web Server Logs Using Apache Hive in Hadoop Framework," pp. 4510–4515, 2015.

[4] K. Jeon, S. Park, S. Chun, and J. Kim, "A Study on the Big Data Log Analysis for Security," vol. 10, no. 1, pp. 13–20, 2016

[5] Miranskyy, A., Hamou-Lhadj, A., Cialini, E., & Larsson, A. (2016). Operational-log analysis for big data systems: Challenges and solutions. IEEE Software, 33(2), 52–59. https://doi.org/10.1109/MS.2016.33

[6] Narkhede, S., & Baraskar, T. (2013). HMR Log Analyzer: Analyze Web Application Logs Over Hadoop MapReduce. International Journal of UbiComp, 4(3), 41–51. https://doi.org/10.5121/iju.2013.4304

[7] S. Adhikari, D. Saraf, M. Revanwar, and N. Ankam, "Analysis of Log Data and Statistics Report Generation Using Hadoop," pp. 4054–4058, 2014.

[8] S. Harish and G. Kavitha, "Statistical Analysis of Web Server Logs Using Apache Hive in Hadoop Framework," pp. 4510–4515, 2015.

[9] I. Technology and R. M. D. Sinhgad, "Real-Time Web Log Analysis and Hadoop for Data Analytics on Large Web Logs," pp. 1827–1832, 2016.
[10] S. Narkhede and T. Baraskar, "HMR LOG ANALYZER: ANALYZE WEB," vol. 4, no. 3, pp. 41–51, 2013.
[11] K. Jeon, S. Park, S. Chun, and J. Kim, "A Study on the Big Data Log Analysis for Security," vol. 10, no. 1, pp. 13–20, 2016.
[12] S. Thosar, A. Mane, S. Raykar, P. Khude, and S. Guru, "Survey on Log Analysis and Management," vol. 3, no. 6, pp. 9–14, 2015.
[13] C. Sharma and A. Jhapate, "A Survey: Analytics of Web Log File through Map Reduce and Hadoop," vol. 2, no. 5, pp. 133–138, 2016.
[14] P. G. Chiddarwar, S. Chhajed, S. Deshmukh, P. Dongre, and B. Nile, "Big Data Analysis of Fetched Logs using Hadoop Framework," vol. 6, no. 3, pp. 706–707, 2017.
[15] L. T. Souza and U. R. Girish, "Error Log Analytics using Big Data and MapReduce," vol. 6, no. 3, pp. 2364–2367, 2015.
[16] B. Gupta, "Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data," vol. 5, no. 3, pp. 3867–3870, 2014.
[17] G. K. Jangla and D. A. Amne, "Development of an Intrusion Detection System based on Big Data for Detecting Unknown Attacks," vol. 4, no. 12, pp. 229–232, 2015.

★★★