ANALYZING LOG IN BIG DATA ENVIRONMENT: A REVIEW

Marlina Abdul Latib, Saiful Adli Ismail, Haslina Md Sarkan and Rasimah Che Mohd Yusoff
Advanced Informatics School, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
E-Mail: marlina.latib@gmail.com

ABSTRACT

Log Analysis is a crucial process in most system and network activities where log data is used for various reasons such as for performance monitoring, security auditing or even for reporting and profiling. However, as years passed by, the volume of log data increases along with the size of the system as well as the number of users involved. Traditional or existing log analyzer tools are not able to handle the massive amount of data. Therefore, Big Data is the solution to overcome this issue. The main purpose of this paper is to present a review of log file analysis in Big Data environment based on previous research works. This paper also highlights the characteristics of Big Data as well as Hadoop Framework that has been widely used as Big Data application. Results from the papers reviewed shows that majority 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 other tools and algorithms together with the Hadoop Framework for analysis purposes. The findings of this paper will provide a comprehensible review of Hadoop usage performance in analyzing different types of log files and recommend understandable results for end users to use in future work.

Keywords: big data, big data characteristics, data preprocessing, Hadoop, HDFS, log analysis, log files, MapReduce

INTRODUCTION

Since decades ago, log data has been playing an important role in computer system. There are various types of log that record different kinds of activities for computer system, applications, network traffic or even web servers. Every details of the log data are crucial in determining the status and condition of a running system. Therefore, log data is one of the main sources in monitoring and analyzing diverse systems and there are many tools designed specifically for analyzing them.

As years passed by, the volume of log data increases along with the size of the systems as well as the number of users involved. This is when data turned into Big Data and more big data problem arises in terms of volume, velocity, and/or variety that exceeds the abilities of most current technology (Joshi, 2013). Big data is actually referring to the architecture and technologies that are established to capture, store, process and run better quality volumes of data but with less amount of time or in real time. Although the massive volume of log can contribute far more comprehensive and valuable information, analyzing them is definitely a great challenge. In the traditional approach, the available data is processed using a powerful computer. However, there is usually a limit to the size of data being processed, as it is not scalable, whereas Big Data expands with great velocity and variety.

If the massive amount of data limits the effectiveness of log file analysis, then Big Data would be the solution. Therefore, this paper will discuss about the log files and the characteristic of Big Data as well as other research works that implement Hadoop framework for analyzing various types of log files. Hadoop is introduced in this paper as an open-source framework that allows distributed processing of massive data sets on clusters of computers which is able to overcome the log analysis problem. The main components of Hadoop, MapReduce and Hadoop Distributed File System (HDFS) are also discussed in this paper.

LOG FILES

Log files are record files that are generated automatically by the source system in nearly all digital devices. They contain huge amount of information which is essential for making business decisions or troubleshooting. For instance, log files will keep information of everything that gets in and out of the web servers. The web servers shall record in the log files the number of clicks, visits or other relevant web users' records which are usually stored in predefined file format (Chen, Mao, and Liu 2014).

Analysis of log files has been very important in resolving many issues. The contribution of the log analysis is categorized into 4 different areas (Oliner, Ganapathi, and Xu, 2012) which are:

a) Performance

Log analysis is used in optimization or debugging process for measuring the system performance. Logs in the case of performance help the administrator to understand how resources of particular system have been used.

b) Security

Logs for security purposes are commonly used to detect breaches or misbehavior and to perform postmortem investigation of security incidents. For instance, intrusion detection needs the reconstruction of sessions from logs in order to detect unauthorized access into a system.

c) Prediction

Logs are also known to be able to produce prediction information. There are predictive analysis tools that use log data to help in marketing strategy, advertisement placement or inventory management.

d) Reporting and profiling

Analyzing logs is also needed in profiling resource utilization, workload or user behavior. For example, logs will record the tasks' characteristics from a cluster's workload in order to profile resource utilization for big data center.

The logging activities can generate huge quantity of logs data causing the log file size to expand up to hundreds of Terabytes per day. However, it will not be an easy task to store and to analyze them when the existing analysis tools are not able to cater such massive amount of data.

Fortunately, by implementing Big Data solution, those logs data can be put to good use as it has the ability to identify large-scale patterns in diagnosing and preventing problems (Sharma and Mangat, 2015).

The following section reviews briefly the definition of Big Data and its characteristics.

BIG DATA

Basically, Big Data is defined as datasets that could not be perceived, acquired, managed and processed within an endurable time by traditional IT and software or hardware tools (Chen, Mao, and Liu, 2014)(Power, 2014). In the industry, Big Data can be defined by two different perspectives (Collins, 2014):

a) Consumers

Big Data is about the use of large datasets from new or various sources in order to provide significant and functional information about how the world works.

b) Producers

It is a technology that is imperative to handle large, diverse datasets, which can be categorized in terms of volume, variety and velocity.

Big Data is not only characterized by three V's: volume, velocity and variety (Sagiroglu and Sinanc, 2013), but also emphasizes two other characteristics which are veracity and value (as illustrated in Figure-1) (Saporito, 2013) (Singh and Kaur, 2014). The following is the description of the related attributes:

a) Volume

Big Data is about 'big' volume when the organizations keep on expanding their petabyte-scale collections of data collected from transaction histories, sensors, click stream or anywhere else (Madden, 2012).

b) Velocity

No matter how big is the data, it must be processed really fast. For instance, the speed of the data processing is needed to determine which advertisement to display to a user web page (Madden, 2012).

c) Variety

Big Data is also referred to as the escalation of new data types from the endless stream of social networks, machine devices, and mobile sources that are consolidated into traditional transactional data types (Lu *et al*., 2014).

d) Veracity

This characteristic is crucial as it involves the truthfulness and understandability of the data where data must be accurately presented (Saporito, 2013).

e) Value

The entire data must be able to give great values for organizations, societies and consumers. (Saporito, 2013).

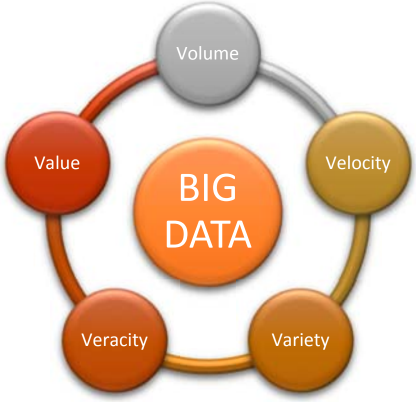


Figure-1. The main characteristics of big data.

In recent research (Gupta, 2015), the Big Data has been suggested to have 2 more characteristic which are:

a) Visualization

The data needs to be readable and easily understood while various optimization algorithms will provide an advantage of providing an optimal review of the data analyzed.

b) Variability

This characteristic allows Big Data to handle uncertainty in data with changing of data helping in prediction of future behavior of the subjects.

Generally, data travels through 4 main different phases as shown in Figure-2. In the Big Data lifecycle, data generation is the first step where data such as Internet data, huge amount of data in terms of searching entries, Internet forum posts, chatting records, and microblog messages, are generated (Chen, Mao, and Liu, 2014).

The second phase of the cycle is data acquisition which includes data collection, data transmission, and data preprocessing. In big data acquisition phase, once the raw data has been collected, for example the log files, there is a need for efficient transmission mechanism to send it to a proper storage management system in order to support different analytical applications.

The preprocessing phase is where activities such as data cleansing or data transformation take place. It is a mandatory and essential step in order to refine and valuate the data. Further details of this phase will be discussed in Data Preprocessing section (Taleb, Dssouli, and Serhani 2015).

Big data storage is related to the storage and management of large-scale datasets in order to achieve reliability and availability of data accessing. In that case, the infrastructure needs to provide information storage service that has reliable storage space as well as a powerful access interface for query and analysis of a large amount of data (Chen, Mao, and Liu, 2014).

The final phase is the data analysis and it is the most important phase in the big data lifecycle. The main purpose of this phase is to extract useful values and provide suggestions or decisions. Through the analysis of datasets in different fields, different levels of potential values can be generated.

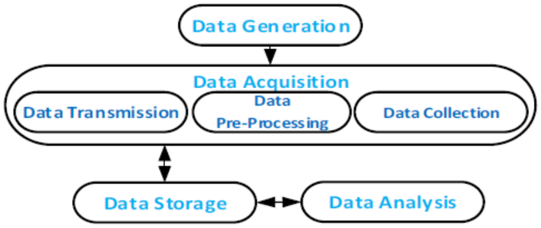


Figure-2. Big data lifecycle (Taleb, Dssouli, and Serhani, 2015).

Due to its specific nature, several applications and solutions have been developed to fulfill the needs of Big Data. One of the widely used application is Hadoop which is an open source framework by Apache. Details of Hadoop framework are discussed in next section.

Hadoop framework

Hadoop is a framework developed by Dog Cutting which is written in Java under the Apache License. It is basically used for analyzing and processing Big Data in a distributed computer environment as well as supporting the running of related applications. Hadoop framework is now becoming solid and stable as well as greater understanding among researchers (Polato *et al*., 2014). Furthermore, Hadoop addresses three main challenges created by Big Data as listed below (Singh and Kaur 2014) (Nandimath *et al*., 2013) (Mohandas and Dhanya, 2013):

a) **Volume:** Hadoop provides a framework to scale out large data sets to address volume of data. It is used in systems where multiple nodes are present, which can process terabytes of data

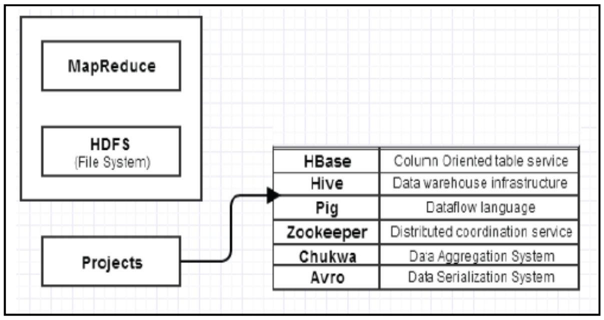
b) **Velocity**: Hadoop is able to handle intense rate of incoming data from very large system. It uses its own file system HDFS which facilitate fast transfer of data which can sustain node failure and avoid system failure as whole.

c) **Variety:** Hadoop supports complex tasks in order to deal with the variety of unstructured data. It uses MapReduce algorithm which breaks down the big data into smaller chunks and performs the operations.

Hadoop framework may vary for each application to another application depending on the needs and expected output. Hadoop is built up of mainly two parts which are HDFS and Map-Reduce Framework (Bhandare, Barua, and Nagare, 2013).

HADOOP ARCHITECTURE

Hadoop consists of two major components which are MapReduce and HDFS as illustrated in Figure-3 (Singh and Kaur, 2014).



HBase	Column Oriented table service
Hive	Data warehouse infrastructure
Pig	Dataflow language
Zookeeper	Distributed coordination service
Chukwa	Data Aggregation System
Avro	Data Serialization System

Figure-3. High level architecture of Hadoop (Singh and Kaur, 2014).

HDFS

HDFS or Hadoop Distributed File System is actually a default storage layer, which is redundantly needed for computations. Based on HDFS architecture, it is more like a master-slave model where a set of Hadoop cluster composes a Namenode as master and many Datanodes as slaves. File system namespace maintenance's tree structure and other information are managed by Namenode while Datanode stores the file system. Datanode is served by several nodes which then return to the Namenode in order to maintain data consistency in providing single directory system and file namespace (Wang *et al*., 2014)(Singh and Kaur, 2014). The architecture of the HDFS is as shown in Figure-4.

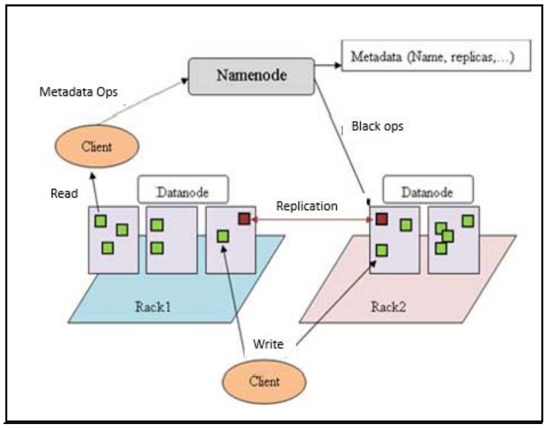


Figure-4. HDFS architecture (Wang *et al.* 2014).

MapReduce

Map/Reduce is a programming paradigm or a software programming model that has been used by Google to process large data set in a distributed fashion on large clusters of hardware in a reliable fault tolerant approach. In MapReduce, task is separated into small parts and distributed to a huge number of nodes for processing (map) and the final answer (reduce) is based on the summarized results. Hadoop uses the Map/Reduce for data processing purposes. Various functions for the processing are written in the form of Hadoop job (Nandimath *et al*., 2013). Figure-5 shows MapReduce processing flow of data. MapReduce library supports application and map operations which can be executed independently (K P, Gouda, and H R, 2015).

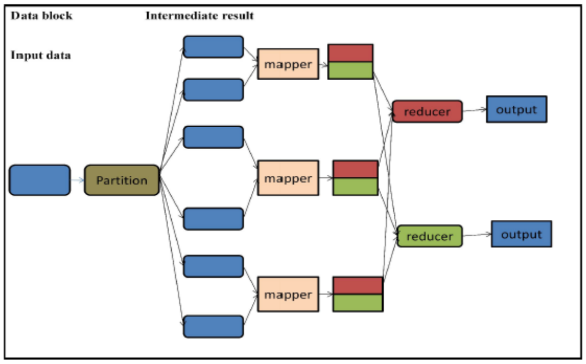


Figure-5. The flow of data in MapReduce processing (K P, Gouda, and H R, 2015).

LOG FILES ANALYSIS

Data preprocessing

When the log data has been collected, it has to go through a preprocessing stage before proceeding to the log analysis stage. This is supported by the data process flow in Figure-6 that shows the ETL (Extract, Transform, and Loading) process is needed in a log analysis flow. The data preprocessing is actually a part of the ETL that transform data to a desired format.

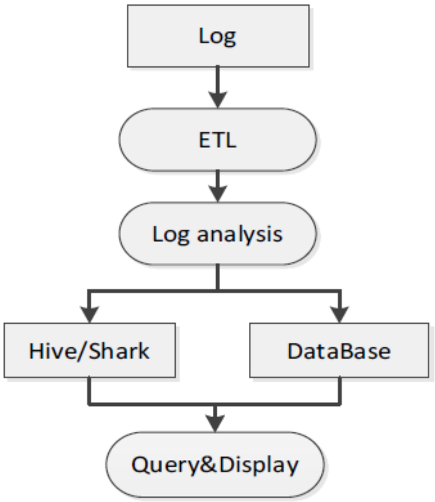


Figure-6. Data process flow (Lin, Wang, and Wu, 2013).

It is crucial for the data to undergo preprocessing operation in order to deal with various imperfections in raw collected data as it may contain noise such as errors, redundancies, outliers and other ambiguous data or missing values. The main operations in data preprocessing are mainly:

- Data cleaning/ data refinement: Handle missing values and noise as well as data inconsistency.
- Data integration: A process of integrating duplicated data.
- Data transformation: The collected data will be converted to the format of the destination data system (Mousannif *et al*., 2014) (Kim and Huh, 2014).

(Li, Yu, and Ryu, 2014) suggested a few steps of preprocessing in the research of web log analysis which involve removing irrelevant attributes or records that have missing valuable data and transforming URLs into code numbers in order to get clean data. Not all of the log records are useful or necessary. Therefore, before the process of web log data analysis is being done, the data cleaning phase needs to be carried out. The data cleaning process involves removing:

- Records that have missing value data, for example, when the execution process is suddenly terminated; the log file record is not completely recorded.
- Illegal records that have exception status numbers for example 400 or 404 which caused by HTTP client errors, bad requests or a request not found.
- Irrelevant records that have no significant URLs. There are some files that are generated automatically when web page is requested, for example .txt, .jpg, .gif or .js extensions.

Therefore, in Log Analysis, the purpose of implementing log preprocessing is to improve the log quality and to increase the results accuracy. The preprocessing phase helps to filter and to organize only appropriate information which is used before applying the MapReduce algorithm so that it may not affect the analysis result.

Related works using Hadoop

The number of research proposing the use of Hadoop Framework in solving Big Data issue is increasing each year. Many of them have been exploring and recommending new log analysis approach using Hadoop in distinctive domains or areas. This paper discussed the related works by other researchers that focuses on log analysis in Big Data environment using Hadoop framework.

1. (Vernekar and Buchade, 2013) proposed the use of Hadoop framework using MapReduce on log analysis for system threats and problem identification. The MapReduce algorithm consists of Map phases and Reduce phases. The log file that needs to be processed will be the input to the Map phase while the output of each map phase will then be given to particular keys. The Reduce function will then provide the final result or log report. The proposed system provides an efficient way of collecting and correlating log in order to identify the system threats and problems. The researchers found that the proposed system has significant improvement in response time, which is achieved with the use of MapReduce.
2. (Wang *et al*., 2014) have designed and implemented and enterprise Weblog analysis systems based on the architecture of Hadoop with HDFS (Hadoop Distributed File System) and MapReduce as well as Pig Latin Language as illustrated in Figure-7. The main purpose of their system is to assist system administrators to quickly capture and analyze data hidden in the massive potential value, thus providing an important basis for business decision. The research that has been carried out showed that the structure of MapReduce program is an effective solution for very large Weblog files in the Hadoop environment. Besides that, the log requirement is easy to analyze using Pig programming language that also gives better performance. The system succeeded in providing AP server traffic statistics that help the system administrators to identify potential problems and predict the future trend.

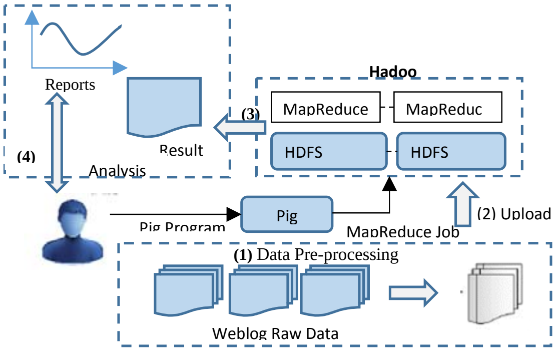


Figure-7. Weblog analysis system flowchart using Hadoop (Wang *et al*., 2014).

3. (Nandimath *et al*., 2013) have proposed a scheme to overcome the problem of analysis of big data using Apache Hadoop in their study. The processing involves four steps which include creating a server of required configuration using Amazon web services, importing data from a database to Hadoop, performing jobs in Hadoop and exporting data back to the database. In step two, data is stored in Mongo DB, which is a NoSQL database. Then, MapReduce is used to perform six Hadoop jobs that are implemented in spring framework. A Hadoop job consists of the mapper and reducer functions. The produced output of data processing in the Hadoop job has to be exported back to the database. The old values in the database have to be updated immediately in order to prevent loss of valuable data. The researchers agreed that the application is able to perform an operation on big data in optimal time and produce an output with minimum utilization of resources.
4. (Therdphapiyanak and Piromsopa, 2013) proposed applying Hadoop for Log Analysis of Apache web servers. They used distributed K-Means clustering algorithm based on Mahout/Hadoop Map-Reduced model to analyze high volume of log files. Their findings showed that the performance was better than a standalone log analyzer as it was capable of supporting a huge size of log. Their method proved to be able to extract a new knowledge for a million entries of logs as it cannot be obtained without the scalability of Hadoop and the proposed analysis.
5. (Hingave and Ingle, 2015) proposed a log analyzer with the combination of Hadoop and MapReduce paradigm

using NASA’s web log file of size 77MB of 445,454 records. They conducted an experiment to do a comparison between MySQL (RDBMS) and Hadoop in analyzing the users’ activity of the web. The results obtained, showed that the proposed log analyzer helped to improve response time as the time required for ETL process and analysis using Hadoop is approximate 20 times less than the MySQL.

6. (Yang *et al*., 2013) proposed an analysis of system for flow log which targeted on the network traffic traces in China to overcome the expanding size of the flow logs which have increased to 870GB per day for single city. Hadoop has been proposed as a solution to make the analysis faster and it is also able to analyze larger dataset. The system uses HDFS for the logs storage and MapReduce Framework for analysis job and also for creating their own script called Log-QL. The results of the experiment show that the new system enabled them to analyze TBs of data compared to the existing centralized system that can only process up to 10GB of data. However, the system will only perform better as the size of data grows.

7. (Narkhede, Baraskar, and Mukhopadhyay, 2014) applied Hadoop MapReduce programming model for analyzing web log files in cloud computing environment in order to retrieve the hit count for specific web application. The experiment uses HDFS to store the web log file and MapReduce programming model is used to write application for analyzing log file. The log files that have been used contain 100,000 records with each log having different fields of URL, date, hit, age and others. The applied model of using HDFS and MapReduce has given analyzed results in minimal response time. While the performance test results against number of records, the number of nodes in the cluster show that the performance of the system will increase along with the increase in number of nodes.

Table-1. Summarized related research of log analysis using Hadoop.

Authors	Hadoop component	Other tools or algorithm	Type of log analysis	Results of research
(Vernekar and Buchade, 2013)	MapReduce		Analyzing sys log to identify system threats	Improve response time
(Wang *et al*., 2014)	HDFS, MapReduce	Pig language	Weblog analysis	Better performance able to predict trend
(Nandimath *et al*., 2013)	MapReduce, HDFS	MongoDB	Amazon log files	Optimal operation time
(Therdphapiyanak and Piromsopa, 2013)	MapReduce Mahout	K-Means	Apache web server log	Support huge file size
(Hingave and Ingle, 2015)	MapReduce	MySQL (RDBMS)	NASA’s web log files	Improve response time
(Yang *et al*., 2013)	HFDS, MapReduce	Log-QL (script)	Flow logs of China network traffic	Able to analyze larger dataset (TBs) Perform better as the data size grows
(Narkhede, Baraskar, and Mukhopadhya, 2014)	HDFS, MapReduce		Web log files in cloud computing	Minimal response time Perform better as the number of nodes increase

Table-1 summarized the related research works of log analysis using Hadoop Framework. The proposed frameworks involving various types of log analysis proved that Hadoop is able to cater and to 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. In order 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 enable us to successfully minimize the analysis process response time as well as analyzing larger dataset (TBs).

CONCLUSIONS

The bigger in size the log files are, the more useful the information they can furnish. However, processing huge sized log data can be very challenging and it is definitively not an easy task. Therefore, the log files should be analyzed in Big Data environment using the Big Data application such as Hadoop. There were researches done to analyze various types of log file using Big Data solution. The outcome is very significant. Results from the papers reviewed shows that majority 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 other tools and algorithms together with the Hadoop Framework for analysis purposes. The results of the related research works presented in this paper showed that by using Hadoop Framework the response and operation time of the analysis process have been improved. Besides that, the bigger the size of the data used, the better the performance would be. The study findings will give direction to our future work to propose a framework of data visualization in log analysis in order to produce understandable results for end users.

ACKNOWLEDGEMENTS

We would like to thank to Advanced Informatics School Universiti Teknologi Malaysia and Ministry of Higher Education (MoHE) for their financial support in this research study.

REFERENCES

[1] Bhandare, Milind, Kuntal Barua, and Vikas Nagare. 2013. Generic Log Analyzer Using Hadoop Mapreduce Framework. International Journal of Emerging Technology and Advanced Engineering 3 (9): 603-7.

[2] Chen, Min, Shiwen Mao, and Yunhao Liu. 2014. Big Data: A Survey. Mobile Networks and Applications 19 (2): 171-209.

[3] Collins, E. 2014. Intersection of the Cloud and Big Data. IEEE Cloud Computing 1 (1): 84–85.

[4] Gupta, A. 2015. Big Data Analysis Using Computational Intelligence and Hadoop: A Study. In 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 1397–1401.

[5] Hingave, H., and R. Ingle. 2015. An Approach for MapReduce Based Log Analysis Using Hadoop. In 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 1264–68.

[6] Joshi, Sanat. 2013. Big Data. InTech 60 (3): 40–43.

[7] Kim, Yong-Hyun, and Eui-Nam Huh. 2014. A Rule-Based Data Grouping Method for Personalized Log Analysis System in Big Data Computing. In 2014 Fourth International Conference on Innovative Computing Technology (INTECH), 109–14.

[8] K P, Ajay, Dr K. C. Gouda, and Dr Nagesh H R. 2015. A Study for Handling of High-Performance Climate Data Using Hadoop. IJITR 0 (0): 197–202.

[9] Li, Meijing, Xiuming Yu, and Keun Ho Ryu. 2014. MapReduce-Based Web Mining for Prediction of Web-User Navigation. Journal of Information Science 40 (5): 557–67.

[10] Lin, Xiuqin, Peng Wang, and Bin Wu. 2013. Log Analysis in Cloud Computing Environment with Hadoop and Spark. In 2013 5th IEEE International Conference on Broadband Network Multimedia Technology (IC-BNMT), 273-76.

[11] Lu, Rongxing, Hui Zhu, Ximeng Liu, J.K. Liu, and Jun Shao. 2014. Toward Efficient and Privacy-Preserving Computing in Big Data Era. IEEE Network 28 (4): 46–50.

[12] Madden, Sam. 2012. From Databases to Big Data. IEEE Internet Computing 16 (3): 4–6.

[13] Mohandas, M., and P.M. Dhanya. 2013. “An Approach for Log Analysis Based Failure Monitoring in Hadoop Cluster.” In 2013 International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), 861-67.

[14] Mousannif, H., H. Sabah, Y. Douiji, and Y.O. Sayad. 2014. From Big Data to Big Projects: A Step-by-Step Roadmap. In 2014 International Conference on Future Internet of Things and Cloud (FiCloud), 373–78.

[15] Nandimath, J., E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi. 2013. Big Data Analysis Using Apache Hadoop. In IEEE 14th International Conference on Information Reuse and Integration (IRI), 700–703.

[16] Narkhede, S., T. Baraskar, and D. Mukhopadhyay. 2014. Analyzing Web Application Log Files to Find Hit Count through the Utilization of Hadoop MapReduce in Cloud Computing Environment. In 2014 Conference on IT in Business, Industry and Government (CSIBIG), 1-7.

[17] Oliner, Adam, Archana Ganapathi, and Wei Xu. 2012. Advances and Challenges in Log Analysis. Commun. ACM 55 (2): 55–61.

[18] Polato, Ivanilton, Reginaldo Ré, Alfredo Goldman, and Fabio Kon. 2014. A Comprehensive View of Hadoop research-A Systematic Literature Review. Journal of Network and Computer Applications 46 (November): 1-25.

[19] Power, Daniel J. 2014. Using ‘Big Data’ for Analytics and Decision Support. Journal of Decision Systems 23 (2): 222–28.

[20] Sagiroglu, S., and D. Sinanc. 2013. Big Data: A Review. In 2013 International Conference on Collaboration Technologies and Systems (CTS). 42-47.

[21] Saporito, Pat. 2013. The 5 V’s of Big Data. Best’s Review, no. 7 (November): 38.

[22] Sharma, S., and V. Mangat. 2015. Technology and Trends to Handle Big Data: Survey. In 2015 Fifth International Conference on Advanced Computing Communication Technologies (ACCT), 266-71.

[23] Singh, K., and R. Kaur. 2014. Hadoop: Addressing Challenges of Big Data. In Advance Computing Conference (IACC), 2014 IEEE International, 686-89.

[24] Taleb, I., R. Dssouli, and M.A. Serhani. 2015. Big Data Pre-Processing: A Quality Framework. In 2015 IEEE International Congress on Big Data (BigData Congress), 191-98.

[25] Therdphapiyanak, Jakrarin, and Krerk Piromsopa. 2013. Applying Hadoop for Log Analysis toward Distributed IDS. In Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, 3:1–3:6. ICUIMC ’13. New York, NY, USA: ACM.

[26] Vernekar, S.S., and A. Buchade. 2013. MapReduce Based Log File Analysis for System Threats and Problem Identification. In Advance Computing Conference (IACC), 2013 IEEE 3rd International, 831–35.

[27] Wang, Chen Hau, Ching TsorngTsai, Chia Chen Fan, and Shyan Ming Yuan. 2014. A Hadoop Based Weblog Analysis System. In 2014 7th International Conference on Ubi-Media Computing and Workshops (UMEDIA), 72-77.

[28] Yang, Jie, Yanshen Zhang, Shuo Zhang, and Dazhong He. 2013. Mass Flow Logs Analysis System Based on Hadoop. In 2013 5th IEEE International Conference on Broadband Network Multimedia Technology (IC-BNMT), 115-18.