ANALYZING MALWARE LOG FILES FOR INTERNET ACCESS INVESTIGATION USING HADOOP

MOHD SHARUDIN MAT DELI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science (Information Assurance)

Advanced Informatics School
Universiti Teknologi Malaysia

DECEMBER 2017

In dedication to my beloved wife, *SUZIANTI YATIMAN*.
In dedication to my beloved children, *IKHWAN, AFIQ, AMNEEY*.
In dedication to my beloved parents, *MAT DELI, SITI PATIMAH*.
In dedication to all lectures and supportive friends.

ACKNOWLEGEMENT

I would like to express my deepest gratitude to my supervisor Dr Saiful Adli Ismail, for his excellent guidance, caring and patience in making my project successful. I am also very thankful to CICT Management (Director and Deputy Director) who have approved and support my study leave. Also to University Teknologi Malaysia for funding my study and giving me this priceless opportunity.

I also place on record, my sense of gratitude to one and all, who directly or indirectly have lent their hand in completing this project.

ABSTRACT

On the Internet, malicious software (malware) is one of the most serious threats to system security. Major complex issues and problems on any software systems are frequently caused by malware. Malware can infect any computer software that has connection to Internet infrastructure. There are many types of malware and some of the popular malwares are botnet, trojans, viruses, spyware and adware. Internet users with lesser knowledge on the malware threats are susceptible to this issue. To protect and prevent the computer and internet users from exposing themselves towards malware attacks, identifying the attacks through investigating malware log file is an essential step to curb this threat. The log file exposes crucial information in identifying the malware, such as algorithm and functional characteristic, the network interaction between the source and the destination, and type of malware. By nature, the log file size is humongous and requires the investigation process to be executed on faster and stable platform such as big data environment. In this study, the authors had adopted Hadoop, an open source software framework to process and extract the information from the malware log files that obtains from university's security equipment. The Python program was used for data transformation then analysis it in Hadoop simulation environment. The analysis includes assessing reduction of log files size, performance of execution time and data visualization using Microsoft Power BI (Business Intelligence). The results of log processing have reduced 50% of the original log file size, while the total execution time would not increase linearly with the size of the data. The information will be used for further prevention and protection from malware threats in university's network.

ABSTRAK

Dalam persekitaran Internet, perisian berbahaya (*malware*) adalah salah satu ancaman yang paling serius terhadap keselamatan sistem. Kebanyakan isu dan masalah rumit yang berlaku dalam sistem adalah disebabkan oleh serangan *malware*. *Malware* boleh menjangkiti mana-mana komputer atau peranti berkaitan yang membuat capaian ke internet. Terdapat pelbagai jenis perisian / aplikasi yang bertujuan jahat, antaranya adalah *botnet*, *trojan*, virus, *spyware* dan *adware*. Pengguna internet yang tidak mengetahui dan kurang berhati-hati akan mudah diserang dan dijangkiti *malware*. Untuk melindungi dan menghalang pengguna komputer dan internet daripada serangan *malware*, mengenalpasti bentuk serangan melalui penyiasatan fail log *malware* merupakan langkah terbaik dalam mengekang pelbagai ancaman. Fail log mendedahkan maklumat penting dalam mengenalpasti *malware* seperti ciri-ciri utama *malware*, algoritma yang digunakan, sumber rangkaian *malware* dan jenis *malware* yang kerap menyerang pengguna. Secara dasarnya, saiz fail log adalah sangat besar dan memerlukan proses analisis dijalankan pada platform yang lebih cepat dan stabil seperti persekitaran *big data*. Dalam kajian ini, penulis telah menggunakan Hadoop yang merupakan kerangka perisian sumber terbuka untuk memproses dan mengekstrak maklumat daripada fail log *malware* yang diperolehi daripada peralatan keselamatan universiti. Program *Python* telah digunakan untuk transformasi data kemudian menganalisisnya dalam persekitaran simulasi Hadoop. Proses analisa ini termasuk menilai pengurangan saiz log fail, prestasi masa pelaksanaan dan visualisasi data yang menggunakan Microsoft Power BI (*Business Intelligence*). Hasil pemprosesan log telah menurunkan 50% dari saiz file log asal, sementara waktu pelaksanaan total tidak akan meningkat secara linear dengan ukuran data. Maklumat ini akan digunakan untuk pencegahan dan perlindungan lanjut daripada ancaman *malware* dalam rangkaian universiti.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

API	-	Application Program Interface
AV	-	Anti Virus
BIOS	-	Basic Input Output System
CNC	-	Command and Control
CSV	-	Comma Separated Values
DOS	-	Disk Operating System
DDOS	-	Distributed Denial of Service
DNS	-	Domain Name System
FTP	-	File Transfer Protocol
HDD	-	Hard Disk Drive
HDFS	-	Hadoop Distributed File System
HDP	-	Hortonworks Data Platform
HTTP	-	Hypertext Transfer Protocol
ICT	-	Information and Communication Technology
IDS	-	Intrusion Detection System
IPS	-	Intrusion Prevention System
IRC	-	Internet Relay Chat
IT	-	Information Technology
LAN	-	Local Access Network
ODBC	-	Open Database Connectivity
OS	-	Operating System
P2P	-	Peer-to-Peer
RAM	-	Random Access Memory
SMTP	-	Simple Mail Transfer Protocol
SQL	-	Structured Query Language
TAN	-	Transaction Authentication Number
TCP	-	Transmission Control Protocol

UDP	-	User Datagram Protocol
URL	-	Uniform Resource Locator
WAN	-	Wide Area Network
YARN	-	Yet Another Resource Negotiator

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Malware in generally is a software or computer programs that is specifically designed by somebody to infiltrate, gain access or damage computers or systems without the user’s consent. There are various types of malware including viruses, spyware, trojan horse, worms, backdoor, keyloggers, exploit, rootkit, or any type of malicious code that infiltrates a computer. Refer to (Devesa *et al*., 2010), malware is the high priority problem and issue to internet security people and also to security researchers and poses a major threat to the privacy information and data of internet users.

Generally, software code is considered malicious code based on the intent of the writer rather than its actual features. Malware creation is on the rise due to the sheer volume of new types created daily and the lure of money that can be made through organized internet crime. Malware was originally created as experiments, but eventually led to criminals and destruction of targeted machines. Based on (Abraham and Chengalur-Smith, 2010), when malware is activated, it makes various changes in the computer by opening backdoors that enable it to spread to other machines. It also executes defensive strategies in order to remain undetected.

Today, a lot of malware is created for profits and gains. The malware attacks through forced advertising such as adware, stealing sensitive information using spyware, spreading email spam or child pornography in zombie computers, or to extort money through ransomware (Hampton and Baig, 2015). Based on (Divya, 2013), various factors can make computers more vulnerable to malware attacks, including weaknesses in operating system design, majority of users are using same OS on the network or uncontrolled and unmanaged user's permission.

The best protection from malware continues to be the usual advice. Refer to (Dang-Pham and Pittayachawan, 2015), internet users must be careful about what email attachments is opened, be cautious when surfing and stay away from suspicious websites, and install and maintain an updated and quality antivirus or anti-malware program.

1.2 Background of the problem

With the growth and vibrancy of internet-based systems and applications, it has increasing number of internet users worldwide. The large number of user have created a big opportunity for cyber criminals to take advantage of internet systems. Internet-based systems and applications has facilitated the business or services, so many users become more dependent on the online environment. Because of that, we can clearly see a massive growth in malware attacks and cyber-criminal activities across the globe.

Malware is a malicious software that can be used by intention person to control overall computer functions, steal confidential and sensitive data, bypass permission access control, or otherwise cause harm to the victim's computer. Malware is referring to the malicious software, a variety of malicious programs. There is variety of malware and based on (Filiol, 2010), the most common types of malware are bots, viruses, worms, spyware and Trojans.

Malware, also refer to using the abridgement malware, it scans internal and external network resources in order to find specific vulnerabilities to exploit it. There are many ways or method for malware to attack and infect computers. The most common method by clicking pop-up or links and subsequently installing a program or code into computers system. The programs or malicious code will execute actions that the user doesn't anticipate or intend.

Execution can be triggered by several user actions, but the most common trigger is a click, typically on a link or pop-up. When user click the link or pop-up message/advertisement, it automatically downloads a malicious software/application into computer. Criminals or attackers will use and manipulate malware applications and codes to take fully control of a computer and then steal confidential data and valuable information.

Refer to (Jing *et al.*, 2015), the increase in malware attacks have caused security hardware to operate more effectively and strongly. With various types of malware and continuously attacks to achieve its objective, will generate a lot of logs in security protection system such as IPS and IDS in organization. The log size keeps increasing and became a large number in Terabytes and Petabytes.

To have strategic protection and prevention in malware attacking and infection, the security log files have to analyse. In analysing a large files and unstructured data, a big data framework and approach is the right choice and the best (Mishra and Singh, 2016, Verma and Pandey, 2016).

1.3 Problem Statement

Malware activities must be analysed for assessment of damage and further prevention. This analysis can determine exactly what happened, how happened and when happened. But, to analyse malware activities is not straight forward and easy because we must work with raw data (log files) obtained from network equipment such as firewall and Internet Access Management with a certain period of time (weekly, monthly, and yearly). With collecting and combination from certain period and various sources to have full analysis, malware log files become large size and the tools used are not capable of handling high volume of data. With the large log files, we need appropriate and powerful system for analysing big data. Therefore, Hadoop is proposed as a framework to process and analyse malware log files and Power BI as analytical tool used for data visualization.

1.4 Research Question

The main research question of this study is how to produce analysis of malware log files for internet investigation using Hadoop. The specific research questions are:

i. What is study related to malware log analysis in internet access environment?
ii. How to identify the proper technology for analysing malware attack in big data environment?
iii. How the analyse malware log files can improve the method in protection and prevention from malware attack?

1.5 Aim of Research

The aim of this research is to improve the process of analysing malware log files obtained from security equipment in order to provide strategic security analysis results.

1.6 Research Objective

The main objective of this project is to produce analysis of malware log files for Internet investigation using Hadoop. While the specific objectives of this study are:

i. To study the attributes and information related to malware log files analysis for internet access investigation.

ii. To design and develop a Hadoop environment for analysis malware log files.

iii. To evaluate performance of execution time and visualize the overall analysis of malware activities / attacks.

1.7 Scope of the Study

The scopes of the research are:

i. The study has only involved in analysing malware log files in order to trace intruder activities.

ii. The work will be focused on analysing internet access log files of University Technology Malaysia, which are obtained from network equipment.

iii. All types of malware are going to analyse including virus, worm, trojan, botnet, spyware, backdoor and rootkits.

iv. The sample log files used in this study are the current and archived log files that have been used for incident investigation.

v. The study will involve Hadoop implementation and log analysis.

vi. Sample log files will be analysed using the Apache Hadoop framework

vii. The simulation will be run using Hortonworks Sandbox with HDP 2.3.2

viii. The simulation will only involve one node cluster by using a single machine.

ix. The analysis tools will involve Hadoop component and Microsoft Power BI (Business Intelligence).

1.8 Significance of the Study

Generally, this study will be a significant contribution to the intrusion investigation in universities and any other organization in term of security, efficiency and cost effectiveness. This study will also benefit to the university Information and Communication Technology (ICT) incident response team in a way to provide security analysis result that will assist in decision making and future prevention strategy regarding network security especially attack by malware. For future research, this study can give researchers another view and technique of malware identification and analysis in proper framework and environment for analysing.

1.9 Thesis outline

This thesis is divided into 6 chapters and organised as follows:

Chapter 1: In this chapter, will discuss an overview of malware in which the specific problem and issues about malware analysis were identified. Then, the objective of this project is clearly stated with specific research question to support research input. The scopes of this project are clearly mentioned and also the significant outcome of this project.

Chapter 2: In this chapter, will summarize the relevant literature in this research area. Paper review that related to malware detection and classification, analysis of log files, especially for security and firewall log, high volume data analysis and big data environment.

Chapter 3: In this chapter, will highlight the research methodology are used to manage and completing this project. It will describe and explains the study outline and method in conducting the research.

Chapter 4: This chapter will explain the configuration and implementation of the proposed simulation model of malware log analysis for internet access investigation. The overall implementation involved two main process which were log processing and log data analysis.

Chapter 5: In this chapter will describe and presents the results of experimental implementation and also the analysis of the results in the research simulation. The results are presented in the form of tables and several types of visualization with detail explanation.

Chapter 6: In this final chapter, will summarize and conclude the overall research project as well as discuss the research contributions, constraints and future works.

REFERENCES

ABRAHAM, S. & CHENGALUR-SMITH, I. 2010. An overview of social engineering malware: Trends, tactics, and implications. *Technology in Society,* 32, 183-196.

ABUZAID, A., SAUDI, M. M., TAIB, B. M. & ABDULLAH, Z. H. 2013. An Efficient Trojan Horse Classification (ETC). *IJCSI International Journal of Computer Science Issues,* 10.

AHMADI, M., SAMI, A., RAHIMI, H. & YADEGARI, B. 2013. Malware detection by behavioural sequential patterns. *Computer Fraud & Security,* 2013, 11-19.

AHMADVAND, H. & GOUDARZI, M. 2016. Using Data Variety for Efficient Progressive Big Data Processing in Warehouse-Scale Computers. *IEEE Computer Architecture Letters,* PP, 1-1.

AJAY, K., GOUDA, K. & NAGESH, H. 2015. A Study for Handelling of High-Performance Climate Data using Hadoop. *IJITR*, 197-202.

AL-SALIM, A. M., ALI, H. M. M., LAWEY, A. Q., EL-GORASHI, T. & ELMIRGHANI, J. M. H. Greening big data networks: Volume impact. 2016 18th International Conference on Transparent Optical Networks (ICTON), 10-14 July 2016 2016. 1-6.

AMAN, S., CHELMIS, C. & PRASANNA, V. Addressing data veracity in big data applications. 2014 IEEE International Conference on Big Data (Big Data), 27-30 Oct. 2014 2014. 1-3.

AMAN, W. 2014. A framework for analysis and comparison of dynamic malware analysis tools. *arXiv preprint arXiv:1410.2131.*

BAI, J. Feasibility analysis of big log data real time search based on Hbase and ElasticSearch. 2013 Ninth International Conference on Natural Computation (ICNC), 23-25 July 2013 2013. 1166-1170.

BAILEY, M., OBERHEIDE, J., ANDERSEN, J., MAO, Z. M., JAHANIAN, F. & NAZARIO, J. 2007. Automated Classification and Analysis of Internet

Malware. *In:* KRUEGEL, C., LIPPMANN, R. & CLARK, A. (eds.) *Recent Advances in Intrusion Detection: 10th International Symposium, RAID 2007, Gold Goast, Australia, September 5-7, 2007. Proceedings.* Berlin, Heidelberg: Springer Berlin Heidelberg.

BAKSHI, K. Considerations for big data: Architecture and approach. 2012 IEEE Aerospace Conference, 3-10 March 2012 2012. 1-7.

BHANDARE, M. & NAGARE, V. 2013. Generic Log Analyzer Using Hadoop Mapreduce Framework. *International Journal of Emerging Technology and Advanced Engineering (IJETAE),* 3.

BHOSALE, H. S. & GADEKAR, D. P. 2014. A Review Paper on Big Data and Hadoop. *International Journal of Scientific and Research Publications,* 4, 1.

CHANG, J., VENKATASUBRAMANIAN, K. K., WEST, A. G. & LEE, I. 2013. Analyzing and defending against web-based malware. *ACM Computing Surveys (CSUR),* 45, 49.

CHEN, M., MAO, S. & LIU, Y. 2014. Big data: A survey. *Mobile Networks and Applications,* 19, 171-209.

CHEN, Z., ROUSSOPOULOS, M., LIANG, Z., ZHANG, Y., CHEN, Z. & DELIS, A. 2012. Malware characteristics and threats on the internet ecosystem. *Journal of Systems and Software,* 85, 1650-1672.

CHEON, J. & CHOE, T.-Y. 2013. Distributed processing of snort alert log using hadoop. *International Journal of Engineering and Technology,* 5, 2685-2690.

DANG-PHAM, D. & PITTAYACHAWAN, S. 2015. Comparing intention to avoid malware across contexts in a BYOD-enabled Australian university: A Protection Motivation Theory approach. *Computers & Security,* 48, 281-297.

DEBATTISTA, J., LANGE, C., SCERRI, S. & AUER, S. Linked 'Big' Data: Towards a Manifold Increase in Big Data Value and Veracity. 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), 7-10 Dec. 2015 2015. 92-98.

DEVESA, J., SANTOS, I., CANTERO, X., PENYA, Y. K. & BRINGAS, P. G. 2010. Automatic Behaviour-based Analysis and Classification System for Malware Detection. *ICEIS (2),* 2, 395-399.

DIVYA, S. 2013. A Survey on Various Security Threats and Classification of Malware Attacks, Vulnerabilities and DetectionTechniques. *International Journal of Computer Science & Applications (TIJCSA),* 2.

DONG, X. L. & SRIVASTAVA, D. Big data integration. Data Engineering (ICDE), 2013 IEEE 29th International Conference on, 2013. IEEE, 1245-1248.

DUPRÉ, L. & DEMCHENKO, Y. Impact of information security measures on the velocity of big data infrastructures. 2016 International Conference on High Performance Computing & Simulation (HPCS), 18-22 July 2016 2016. 492-500.

EICK, S. G., NELSON, M. C. & SCHMIDT, J. D. 1994. Graphical analysis of computer log files. *Communications of the ACM,* 37, 50-56.

FAN, W. & BIFET, A. 2013. Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter,* 14, 1-5.

FILIOL, E. 2010. Viruses and Malware. *In:* STAVROULAKIS, P. & STAMP, M. (eds.) *Handbook of Information and Communication Security.* Berlin, Heidelberg: Springer Berlin Heidelberg.

GASPARY, L. P., MELCHIORS, C., LOCATELLI, F. E. & DILLENBURG, F. Identification of intrusion scenarios through classification, characterization and analysis of firewall events. 29th Annual IEEE International Conference on Local Computer Networks, 16-18 Nov. 2004 2004. 327-334.

GHAZI, M. R. & GANGODKAR, D. 2015. Hadoop, MapReduce and HDFS: A Developers Perspective. *Procedia Computer Science,* 48, 45-50.

HAMPTON, N. & BAIG, Z. A. 2015. Ransomware: Emergence of the cyber-extortion menace.

HASHEM, I. A. T., YAQOOB, I., ANUAR, N. B., MOKHTAR, S., GANI, A. & KHAN, S. U. 2015. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems,* 47, 98-115.

HELLAL, A. & BEN ROMDHANE, L. 2016. Minimal contrast frequent pattern mining for malware detection. *Computers & Security,* 62, 19-32.

HINGAVE, H. & INGLE, R. An approach for MapReduce based log analysis using Hadoop. 2015 2nd International Conference on Electronics and Communication Systems (ICECS), 26-27 Feb. 2015 2015. 1264-1268.

IDIKA, N. & MATHUR, A. P. 2007. A survey of malware detection techniques. *Purdue University,* 48.

INDRE, I. & LEMNARU, C. Detection and prevention system against cyber attacks and botnet malware for information systems and Internet of Things. 2016

IEEE 12th International Conference on Intelligent Computer Communication and Processing (ICCP), 8-10 Sept. 2016 2016. 175-182.

ISLAM, R., TIAN, R., BATTEN, L. M. & VERSTEEG, S. 2013. Classification of malware based on integrated static and dynamic features. *Journal of Network and Computer Applications,* 36, 646-656.

JING, Y., TINGWEN, L., HAOLIANG, Z., JINQIAO, S. & GUO, L. An automatic approach to extract the formats of network and security log messages. MILCOM 2015 - 2015 IEEE Military Communications Conference, 26-28 Oct. 2015 2015. 1542-1547.

KALIGE, E., BURKEY, D. & DIRECTOR, I. 2012. A case study of Eurograbber: How 36 million euros was stolen via malware. *Versafe (White paper),* 35.

KAMIYA, K., AOKI, K., NAKATA, K., SATO, T., KURAKAMI, H. & TANIKAWA, M. The method of detecting malware-infected hosts analyzing firewall and proxy logs. Information and Telecommunication Technologies (APSITT), 2015 10th Asia-Pacific Symposium on, 2015. IEEE, 1-3.

KANAKER, H. M., SAUDI, M. M. & MARHUSIN, M. F. Detecting worm attacks in cloud computing environment: Proof of concept. 2014 IEEE 5th Control and System Graduate Research Colloquium, 11-12 Aug. 2014 2014. 253-256.

KATAL, A., WAZID, M. & GOUDAR, R. H. Big data: Issues, challenges, tools and Good practices. 2013 Sixth International Conference on Contemporary Computing (IC3), 8-10 Aug. 2013 2013. 404-409.

KHAN, M. A. U. D., UDDIN, M. F. & GUPTA, N. Seven V's of Big Data understanding Big Data to extract value. Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 3-5 April 2014 2014a. 1-5.

KHAN, N., YAQOOB, I., HASHEM, I. A. T., INAYAT, Z., MAHMOUD ALI, W. K., ALAM, M., SHIRAZ, M. & GANI, A. 2014b. Big data: survey, technologies, opportunities, and challenges. *The Scientific World Journal,* 2014.

KOMAR, M., SACHENKO, A., KOCHAN, V. & SKUMIN, T. Increasing the resistance of computer systems towards virus attacks. 2016 IEEE 36th International Conference on Electronics and Nanotechnology (ELNANO), 19-21 April 2016 2016. 388-390.

KONIARIS, I., PAPADIMITRIOU, G., NICOPOLITIDIS, P. & OBAIDAT, M. Honeypots deployment for the analysis and visualization of malware activity and malicious connections. 2014 IEEE International Conference on Communications (ICC), 10-14 June 2014 2014. 1819-1824.

KUNG, S.-Y. Visualization of big data. Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on, 2015. IEEE, 447-448.

LABOSHIN, L. U., LUKASHIN, A. A. & ZABOROVSKY, V. S. 2017. The Big Data Approach to Collecting and Analyzing Traffic Data in Large Scale Networks. *Procedia Computer Science,* 103**,** 536-542.

LAKAVATH, S. & RAMLAL NAIK, L. 2014. A Big Data Hadoop Architecture for Online Analysis. *IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN***,** 2249-9555.

LI, Z. & OPREA, A. Operational Security Log Analytics for Enterprise Breach Detection. 2016 IEEE Cybersecurity Development (SecDev), 3-4 Nov. 2016 2016. 15-22.

LUBNA, K., CYIAC, R. & KARUN, A. K. Firewall log analysis and dynamic rule re-ordering in firewall policy anomaly management framework. Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on, 2013. IEEE, 853-856.

LUO, Y., LUO, S., GUAN, J. & ZHOU, S. 2013. A RAMCloud Storage System based on HDFS: Architecture, implementation and evaluation. *Journal of Systems and Software,* 86**,** 744-750.

MANGIALARDO, R. J. & DUARTE, J. C. 2015. Integrating Static and Dynamic Malware Analysis Using Machine Learning. *IEEE Latin America Transactions,* 13**,** 3080-3087.

MATSUMOTO, S., SATO, A., SHINJO, Y., NAKAI, H., ITANO, K., SHOMURA, Y. & YOSHIDA, K. A method for analyzing network traffic using cardinality information in firewall logs. Applications and the Internet (SAINT), 2010 10th IEEE/IPSJ International Symposium on, 2010. IEEE, 241-244.

MISHRA, A. D. & SINGH, Y. B. Big data analytics for security and privacy challenges. 2016 International Conference on Computing, Communication and Automation (ICCCA), 29-30 April 2016 2016. 50-53.

MOHANDAS, M. & DHANYA, P. 2013. An exploratory survey of Hadoop log analysis tools. *International Journal of Computer Applications,* 75.

MOSER, A., KRUEGEL, C. & KIRDA, E. Exploring Multiple Execution Paths for Malware Analysis. 2007 IEEE Symposium on Security and Privacy (SP '07), 20-23 May 2007 2007. 231-245.

OKANE, P., SEZER, S. & MCLAUGHLIN, K. 2011. Obfuscation: The Hidden Malware. *IEEE Security & Privacy,* 9, 41-47.

PATIL, T. R. & SHEREKAR, S. 2013. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications,* 6, 256-261.

PODILE, A., KEERTHI, G. & PENDYALA, K. S. 2015. Digital Forensic analysis of malware infected machine–Case study. *International Journal of Scientific & Technology Research,* 4.

POLATO, I., RÉ, R., GOLDMAN, A. & KON, F. 2014. A comprehensive view of Hadoop research—A systematic literature review. *Journal of Network and Computer Applications,* 46, 1-25.

PROVOS, N., MCNAMEE, D., MAVROMMATIS, P., WANG, K. & MODADUGU, N. 2007. The Ghost in the Browser: Analysis of Web-based Malware. *HotBots,* 7, 4-4.

RAMANI, R. G., KUMAR, S. S. & JACOB, S. G. Rootkit (malicious code) prediction through data mining methods and techniques. 2013 IEEE International Conference on Computational Intelligence and Computing Research, 26-28 Dec. 2013 2013. 1-5.

REDDY, K. S., REDDY, M. K. & SITARAMULU, V. An effective data preprocessing method for Web Usage Mining. Information Communication and Embedded Systems (ICICES), 2013 International Conference on, 2013. IEEE, 7-10.

SAGIROGLU, S. & SINANC, D. Big data: A review. Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013. IEEE, 42-47.

SAMAK, T., GUNTER, D. & HENDRIX, V. Scalable analysis of network measurements with Hadoop and Pig. 2012 IEEE Network Operations and Management Symposium, 16-20 April 2012 2012. 1254-1259.

SARANYA, R., KANNAN, S. S. & PRATHAP, N. A survey for restricting the DDOS traffic flooding and worm attacks in Internet. 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 29-31 Oct. 2015 2015. 251-256.

SHETA, M. A., ZAKI, M., HADAD, K. A. E. S. E. & H. A, M. Anti-spyware Security Design Patterns. 2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 21-23 July 2016 2016. 465-470.

SHIJO, P. V. & SALIM, A. 2015. Integrated Static and Dynamic Analysis for Malware Detection. *Procedia Computer Science,* 46, 804-811.

SHU, X., SMIY, J., YAO, D. D. & LIN, H. Massive distributed and parallel log analysis for organizational security. 2013 IEEE Globecom Workshops (GC Wkshps), 9-13 Dec. 2013 2013. 194-199.

SINGH, N. K., TOMAR, D. S. & ROY, B. N. 2010. An approach to understand the end user behavior through log analysis. *International Journal of Computer Applications,* 5, 27-34.

SIWOON, S., MYEONG-SEON, G. & MOON, Y. S. Anomaly detection for big log data using a Hadoop ecosystem. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 13-16 Feb. 2017 2017. 377-380.

STEWART, J. 2006. Behavioural malware analysis using Sandnets. *Computer Fraud & Security,* 2006, 4-6.

TAZAKI, H., OKADA, K., SEKIYA, Y. & KADOBAYASHI, Y. MATATABI: Multi-layer Threat Analysis Platform with Hadoop. 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), 11-11 Sept. 2014 2014. 75-82.

TIAN, R., ISLAM, R., BATTEN, L. & VERSTEEG, S. Differentiating malware from cleanware using behavioural analysis. 2010 5th International Conference on Malicious and Unwanted Software, 19-20 Oct. 2010 2010. 23-30.

UZUNKAYA, C., ENSARI, T. & KAVURUCU, Y. 2015. Hadoop Ecosystem and Its Analysis on Tweets. *Procedia - Social and Behavioral Sciences,* 195, 1890-1897.

VAYSTIKH, A., POLANSKY, R., SAKLIKAR, S. D. & LIPTZ, L. 2013. Malware detection using risk analysis based on file system and network activity. Google Patents.

VENKATESAN, N. J., EARL, K. & DONG RYEOL, S. PoN: Open source solution for real-time data analysis. 2016 Third International Conference on Digital

Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 6-8 July 2016 2016. 313-318.

VENUGOPALAN, V., PATTERSON, C. D. & SHILA, D. M. Detecting and thwarting hardware trojan attacks in cyber-physical systems. 2016 IEEE Conference on Communications and Network Security (CNS), 17-19 Oct. 2016 2016. 421-425.

VERMA, C. & PANDEY, R. Big Data representation for grade analysis through Hadoop framework. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 14-15 Jan. 2016 2016. 312-315.

WAZID, M., KATAL, A., GOUDAR, R. H., SINGH, D. P., TYAGI, A., SHARMA, R. & BHAKUNI, P. A framework for detection and prevention of novel keylogger spyware attacks. 2013 7th International Conference on Intelligent Systems and Control (ISCO), 4-5 Jan. 2013 2013. 433-438.

WEN, W., XIAOFENG, W., HUABIAO, L. & JINSHU, S. Automatic signature analysis and generation for large-scale network malware. IET International Conference on Information Science and Control Engineering 2012 (ICISCE 2012), 7-9 Dec. 2012 2012. 1-5.

ZHAO, D., TRAORE, I., SAYED, B., LU, W., SAAD, S., GHORBANI, A. & GARANT, D. 2013. Botnet detection based on traffic behavior analysis and flow intervals. *Computers & Security,* 39, Part A, 2-16.