A NOVEL DISTRIBUTED INTRUSION DETECTION FRAMEWORK FOR NETWORK ANALYSIS

Rashmi MR[1], M Sudheep Elayidom[2], R VijayaKumar[3]

[1] *M.Tech Scholar, SOCS, M G University Kottayam, Kerala, India,***rashmicusat@gmail.com**
[2] *Associate Professor, CUSAT, Kochi, India,* **sudheepelayidom@hotmail.com**
[3] *Professor, SOCS, MG University Kottayam, Kerala, India,* **vijayakumar@mgu.ac.in**

Abstract

Computer networks are used to transfer information between different types of computer devices. Due to rapid development in internet technologies, network users and communication increases day by day. Hence there is a need for huge data analysis, but a currently available tool has been facing a bottleneck. The volume of data along with the speed it generates makes it difficult for the current available tools to handle big data. To overcome this situation, big data packet analysis can be performed through a cloud computing platform for distributed storage (HDFS) and distributed processing (map reduce). However, with the extensive use of cloud computing, security issues arise. With increase of networks, security methods also need to be increased day by day. Hence, intrusion detection system (IDS) are essential components in secure network environment monitors network traffic and allows early detection attacks and alerts the system. Snort is most commonly used IDS available under GPL, which allows pattern search. Hence, there is an urgent need to intelligent intrusion detection systems (IDSs) to detect intrusions automatically. The functionality of Snort IDS can be extended by integrating anomaly preprocessor to detect new attacks. This paper provides a novel distributed Intrusion detection framework for network analysis using snort and Hadoop.

Key Words: *IDS, Snort, Big data, Hadoop, HDFS, Map Reduce and Anomaly preprocessor*

---***---

1. INTRODUCTION

When technology advances, computer network usage as well as network crime increased.in other words, any attack can be arrive from any node. Hence it is better to detect early and subsequent actions should be taken to avoid further attacks.

An intrusion detection system (IDS) is important components in a secure network environment that monitors network traffic that allows early identification and alerts the system. IDS comprise of three fundamental components Source, Analysis, and action. Information source can be network based, host based or hybrid [3]. IDS comes under two category misuse detection /signature based IDS (pre-established one) and anomaly detection (newly launched one).Misuse detection is based on knowledge of patterns related with known attacks provided by human. Anomaly detector generates profiles that represent normal data and any deviation from these profiles can be considered as attack. Statistical methods, expert system are some of the methods for intrusion detection based on Anomaly detection. It generates alert logs when a possible intrusion occurs in the system [2][4].

2. BACKGROUNDS

2.1 Intrusion Detection Systems

Security attacks mainly falls under 2 classes: Active and Passive. Attackers are hidden in case of active attacks and it taps the communication link to collect data. Passive attacks can be grouped into eavesdropping, node malfunctioning, node destruction and traffic analysis types. In active attacks, affects the operations in the attacked network and can be detected. Active attacks can be grouped into Denial-of-Service (DoS), jamming,hole attacks (black hole, wormhole, sinkhole, etc.), flooding and Sybil types. Solutions to security attacks involve three main components Prevention (defense against attack),Detection (being aware of the attack that is present) and Mitigation (reacting to the attack)[2].

Intrusion Detection Systems (IDSs) provide some or all of the following data to the other systems: intruder identity, intruder location, time, intrusion type, layer where it effected. These data are very useful in mitigating .Hence, intrusion detection systems are very important for network security[2][4][3].

2.2 Snort

Snort is a free and open source network intrusion detection and prevention system created by Martin Roesch in 1998. It runs under three modes: a sniffing mode, a logging mode and IDS mode. Sniffer modes read the network packets and display them. Packet logger mode logs the network packets to the disk. Network IDS is the most difficult mode. It monitors network traffic and compare with a rule dataset defined by the user and then perform a corresponding action [5]. In intrusion detection mode, Snort does not capture packet as it does in the network sniffer mode. If a packet matches a rule, then only it is logged or an alert is generated, else no log entry is created. A Snort-based IDS consists of the following major components:

[1] Packet Decoder: The packet decoder collects packets and sent to the detection engine.

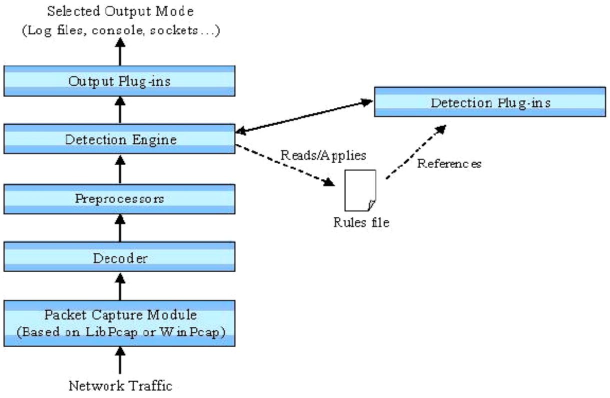

Fig- 1: Components of Snort

[1] Preprocessors or Input Plug-ins: Preprocessors are plugins used with Snort to modify packets before detection engine.

[2] Detection Engine: It is the most important part of Snort. It detect intrusion exists in any packets. It employs Snort rules for this purpose. If a packet matches any rule, appropriate action is taken; otherwise the packet is dropped. Appropriate actions may be logging the packet or generating alerts.

[3] Logging and Alerting System: It generates alert and log messages

[4] Output Modules: Output modules process alerts and logs[5].

Snort Rules:

The Snort programs are coded and implemented by Source fire Inc. Snort system is installed in the Linux environment. Then configure rules in snort.conf configuration files. The rule syntax in Snort is well defined. Each rule consists of two parts: rule Header composed of the 5 tuples such as Action, Protocol Address, Port Direction, Address Port, Keyword separator, argument, delimiter & rule options. These rules consist of rules-head and rules-body, rule-head is a rule action such as alert agreement, source IP, source port, destination IP, and destination port. Rule bodies have the alarm information and characteristics of string. The rules are of the form ``cond -> action", where action specifies the action to be taken on a packet that matches the condition cond[6]. Rules are usually placed in a configuration file snort.conf[5].

Snort intrusion detection mode:

In the Network Intrusion Detection (NID) mode, it generates alerts when a captured packet matches a rule. Snort can send alerts in many modes. These modes are configurable through the command line in snort.conf file. It reads its configuration file /etc/syslog.conf where the location of these log files is configured. The usual location of syslog files is /var/log directory [6].

2.3 Hadoop

Hadoop is for Big Data Analysis. Hadoop is for analysing peta bytes of data in a very short span of time. Hadoop is basically a framework for running applications on large clusters .It enables thousands of nodes to work together in parallel for doing a single job. As the number of nodes increases, the time taken to process the data decreases[8][13].

The two important features of Hadoop are Distributed Storage and Distributed processing. Distributed Storage is given by Hadoop Distributed File System (HDFS) and Distributed Processing is done by the concept known as Map Reduce. A small Hadoop cluster will include a single master and multiple worker nodes (slaves).The master node consists of a Job Tracker, Task Tracker, Name Node and Data Node. A slave or worker node acts as both a Data Node and Task Tracker. Job Tracker and Task Trackers are responsible for doing the map reduce jobs. Name nodes and Data nodes are the part of the distributed file system[7].

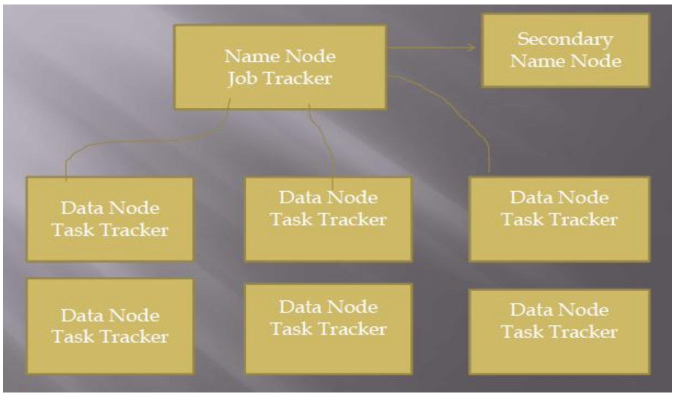

Fig-2: Hadoop Master Slave Architecture

HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Files are divided into large blocks and are distributed across the cluster. Typical block size is 64 MB. Block sizes can be changed. To handle failures, blocks will be replicated by appropriate replication factor[7].

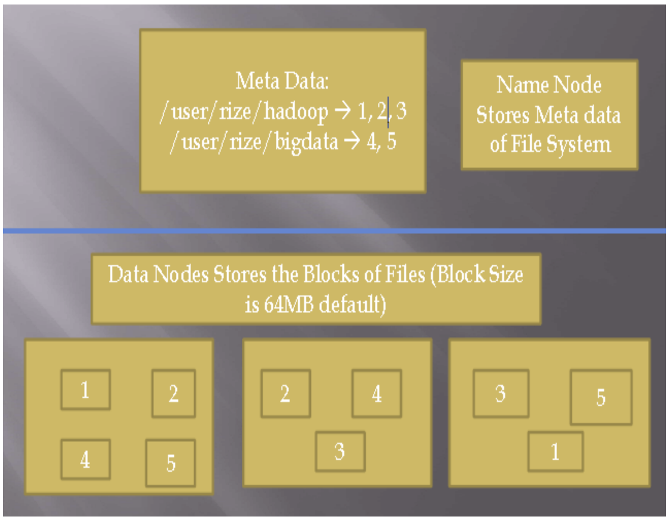
Meta Data:
/user/rize/hadoop → 1, 2, 3
/user/rize/bigdata → 4, 5
Name Node
Stores Meta data
of File System
Data Nodes Stores the Blocks of Files (Block Size
is 64MB default)
1
2
4
5
2
4
3
3
5
1

Fig- 3: HDFS

HDFS cluster is formed by a cluster of Data Nodes. A Name Node performs various file system operation such close, open rename etc. Name Node also manages block replication. Data Nodes run operations such as read and write that file system clients require. A file is divided into more than a block that is stored in a Data Node and HDFS determines mapping between Data Nodes and blocks. Map Reduce is a software framework for writing applications which process vast amounts of data in-parallel on large clusters having thousands of nodes in a reliable, fault-tolerant manner[7][10].

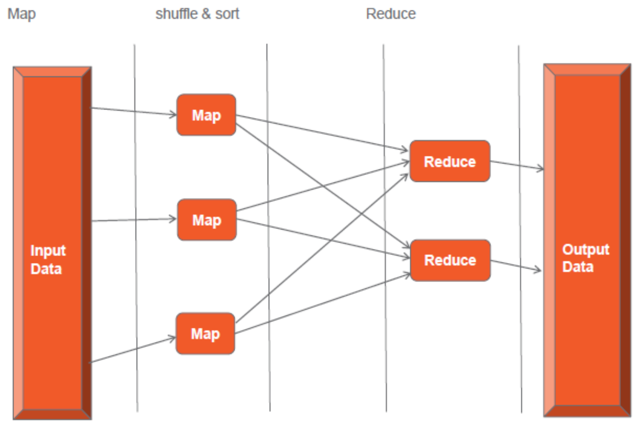

Fig- 4: Map Reduce

The Map Reduce engine which is placed above the file system consists of one Job Tracker. Client applications submit Map Reduce jobs to this Job Tracker. The Job Tracker pushes work out to available Task Tracker[10]. One of the most highlighted feature of Hadoop is that is rack-aware. With rack-aware file system, the Job Tracker knows which node contains the data, and which other machines are nearby. If node fails, work is given to nodes in the same rack. This reduces network traffic on the main backbone network[10].

In the Map Reduce programming model used in Hadoop, the computation takes a set of input key/value pairs, and produces a set of out-put key/value pairs. Map and Reduce are two basic functions in the Map Reduce computation[10].

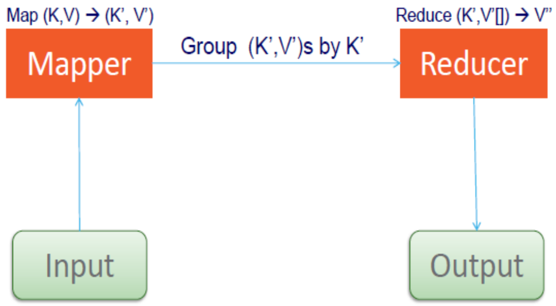

Fig- 5: Map Reduce Key Value

3 EXISTING SYSTEM CHALLENGES

3.1 Problem Definition

The main aim of any IDS system is to effectively analyze all packets passing throw the network without any packet drops. The Internet traffic is rapidly increasing due to high-speed and will continue to grow. The main limitation is to deal with this huge amount of traffic data. However, there is no any analysis tool that can afford this much amount of data at once.

With the rapid development of the Internet, the network security issues become more and more serious. Processor speed growth has not kept up with the growth of the network traffic and the rapid growth of network bandwidth. So it is difficult to complete the analysis of the flow of hundreds of megabytes by intrusion detection equipment, not to mention on Gigabit network traffic. So network traffic is stored and process in a cloud computing platform Hadoop. There are several research works being done to impart better performance to Snort & Hadoop [1][9][11].

3.2 Problem Solution

Hadoop for its scalability in storage and computing power is a suitable platform for Internet traffic measurement and analysis but brings about several research issues.[1][9] It analyses the network data packet which contains intrusion features with a high degree of regularity and extracts abnormal flow characteristics, so it can effectively improve the efficiency of intrusion detection. Finally, combine the snort detection engine with the distributed programming Map Reduce model, and make it suitable for concurrent processing which can effectively respond to the massive data packets in the high-speed network [11].Artificial intelligence can be integrated to Snort to enhance its feature. This improved Snort was able to differentiate between normal traffic and malicious one.

3.3 Proposed System

In this work, the intrusion detection system Snort is made use of. Snort is basically a Signature based IDS (detect only known attack) which typically works on set of rules. The incoming packets capture through snort packet logger mode and stored as log files in local disk.in real life situation these log files are of huge amount.so hadoop cloud computing platform is used for big data analysis. After analysis these packets are compared with the set of rules. If any of the packets matches with the set of rules, actions specified in the corresponding rules are performed [6]. The main disadvantage of Signature based intrusion detection system is that it is unable to identify unknown attacks. This kind of intrusion detection systems work with rules. Therefore those attacks those are not present as rule cannot be detected. Such kind of attacks are identified by Anomaly based Intrusion Detection Systems. Anomaly based ids, as the name suggests typically works on any anomaly .i.e. any deviation from the normal pattern is treated as anomaly and such anomalies will be classified as attacks. Thus it will be able to identify newly launched ones also. Anomaly preprocessor is added to snort that extends functionality of snort. The main disadvantage of this system is the generation of large number of false attacks. However, high volumes of network traffic requiring new approaches that are able to handle huge volume of log and packet analysis [1][9] .

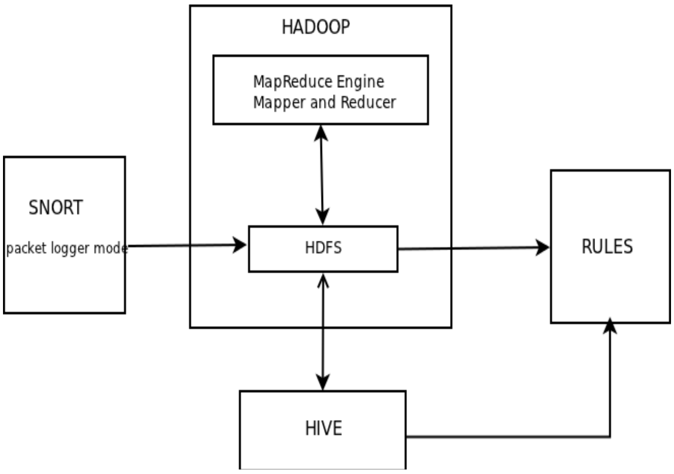

Fig-6: Proposed Architecture

Hadoop, an open-source computing platform of Map Reduce and a distributed file system, has become a popular infrastructure for massive data analytics because it facilitates scalable data processing and storage services on a distributed computing system consisting of commodity hardware. The proposed architecture is able to efficiently handle large volumes of collected data and consequent high processing loads using Hadoop, Map Reduce and cloud computing infrastructure [10]. The main focus of the paper is to enhance the throughput and scalability of Log analysis [13]. In this work, the packets captured by Snort are analyzed by the Grid computing framework Hadoop, which is used for Big Data Analysis. In this work, map reduces[12] analyzed data stored at HDFS to generate the count of the number of packets between any pair of nodes. For those ip addresses that generate large number of packets, Snort rules will be generated so that when the number of packets from a particular source exceeds a number, the node will generate alerts to other nodes since there is a possibility of attack. Hadoop analyses the set of packets and corresponding customized Snort rules are generated if number of packets exceeds a count than the normal one. By employing information provided by IDS, it is possible to apply appropriate countermeasures and mitigate attacks that would otherwise seriously undermine network systems.

4 SYSTEM DESIGN

4.1 Data Flow Diagram

A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system

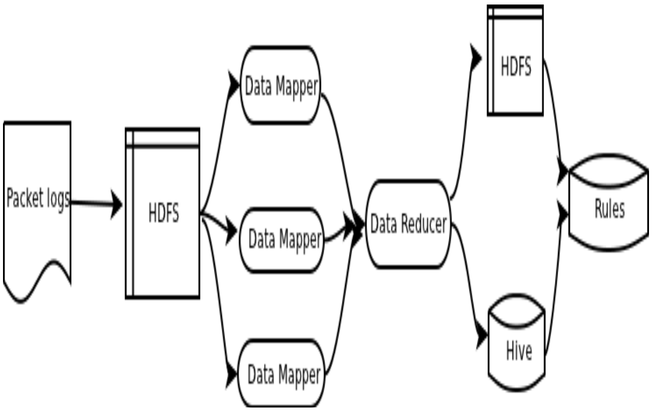

Fig-7: Overall Dataflow

5. IMPLEMENTATION

5.1 Log File Creation

Hping3 is used to generate packets of different format that simulate flooding attacks. Hping3 commands used are as follows: sudo hping3 --icmp --flood 192.168.1.1.The incoming packets are capture through snort packet logger mode and stored as log files in local disk. Different snort commands are specified below: sudo snort -l /var/log/snort -c /etc/snort/snort.conf -i wlan0(snort in IDS mode);(snort in logger mode is as follow:sudo snort -ved -h 192.168.1.0/24 -l /home/rashmi;) ls snort*;(to list log file generated)command to convert to human readable form is sudo snort -r snort.log.1427268044.The log files where then converted into suitable readable format using snort command or wireshark. Out of them source ip, destination ip, source port, destination port, protocol and count where extracted and converted to csv format using wireshark [6].

5.2 Packet Analysis using Hadoop

Snort rules generated after analysis is very much efficient in detecting many attacks. To enhance the throughput and scalability of log analysis cloud computing infrastructure Hadoop is used[1][9][11].Hadoop cdh 4 is installed and checked for wordcount program first. Default folders are set for Hadoop input and output at HDFS. Before running Hadoop Map reduce program written in java programming language, log file generated are loaded into input folder in csv format. Map reduce program use key as source ip, destination ip and protocol; value obtained will be count of packets from source.

Hive a data ware house is added to Hadoop for more user friendliness.it uses HiveQL similar to Mysql .Some of Hive query is mentioned below:

5.3 Snort Detection Engine

The new version of Snort uses an improved detection engine for matching signatures[5]. Each incoming packet is compared with snort rule database, if matches perform action specified in rule. For those ip addresses that generate large number of packets, Snort rules will be generated so that when the number of packets from a particular source exceeds a number, the node will generate alerts to other nodes since there is a possibility of attack[6].

5.4 Snort with Anomaly Preprocessor

Anomaly based IDS collects incoming packets and a normal profile is generated based on baseline that include variables such as host memory or CPU usage, network packet types, and packet quantities. Anomaly-based intrusion detection triggers an alarm on the IDS when some type of unusual behavior occurs on your network. Different methods such as KNN classifier, naïve Bayes classifier, random forest classifier and neural networks are used for this.

The advantage of the anomaly-based approach is that the IDS can detect new types of attacks because it is looking for abnormal activity of any type. Rule corresponding to new attack is entered in snort database for further reference[6].

6. RESULTS

Map Reduce job was done on Hadoop cluster with various file sizes and the CPU times consumed for each file size is noted. The number of reduce tasks in all cases is one. As shown in results the analysis process does not take much time to perform the task.

Log File Size	CPU time Spent
500 MB	54.123 s
800 MB	109.013 s
2 GB	150.231 s
2.7 GB	285.170 s

Table -1: Running With Various File Size

In Hive, the time for retrieving data is very less. During implementation, the results of analysis were loaded to Hive. The queries which included joins and aggregate functions are carried out as Map Reduce jobs. Therefore it can be concluded that managing data, with certain format is very much efficient with Hive.

Hive query	CPU time spent
Show tables	8.914 s
Create table	0.582 s
Load data	2.014 s
Select * from table	14.707 s
Select * from table for cond	19.547 s
Drop table	3.222 s

Table -2: Running with Various hive query

Anomaly based preprocessor with different methods are added with snort to detect new anomalies and corresponding rule is added to snort rule database for further reference. The performance evaluations of these methods are done using KDD CUP 99 dataset. Anomaly detection is done using naïve Bayes classification algorithm and K-nearest neighbor classifier with KDD 99 dataset with 494021 rows. Naive Bayes classifier takes less time for prediction compared to K-nearest neighbor algorithm for prediction .But the main disadvantage of Naïve Bayes Classifier is less accuracy compared to other methods such as K-nearest neighbor classifier, Random forest Classifier and neural network classifiers.

7. CONCLUSIONS AND FUTURE SCOPE

Security is the most crucial issue that is taking place. Large amount of data has to be analyzed, for finding anomaly. A user can customize its own rules even without analyzing. But it will generate a lot of false alarms i.e. a node which is not an attacker may be treated as an attack. Snort rules generated after analysis is very much efficient in detecting many attacks. In future the work can be extended to find more attacks. For those ip addresses that send large number of packets than normal ones, snort rules were generated. The rules where generated by adding options for event filtering in such a way that the alerts will be generated only if number of packets from the particular source exceeds a particular number. This is to avoid the number of false alarms.

The importance of Internet traffic analysis, measurement, and classification is demanding as the traffic data grows. Still there are multiple issues to be considered like real time traffic analysis; the research is going on the same. In future machine learning algorithms integrated with Hadoop will play an important role in traffic classification.AI is integrated into Snort preprocessor plug-in, which makes Snort IDS more intelligent to detect new or variant network attacks. Future works includes, as for detection engineer of Snort IDS, some evolutionary algorithms such as genetic algorithm (GAs) or immune algorithms (IAs) approaches can be combined with it.

Further techniques can be developed to minimize false alarms. Study can be further extended to compare Snort with more parameters such as false alarms, maintenance, firewall etc. Even though our architecture still use off-line analysis tool Hadoop, future work will consist in implementing the same system with on-line analysis big data framework such as Spark Streaming or Storm. The future work may implement IPS to detect and prevent the threats against the virtualized environment.

ACKNOWLEDGEMENT

We would like to thank all friends, faculties and students of CUSAT and MG University who helped us in completing this work.

REFERENCES

[1] Yeonhee Lee and Youngseok Lee, “Toward scalable internet traffic measurement and analysis with Hadoop”. ACMSIGCOMM Comput. Commun. Rev. 43 Issue 1, Pages 5 – 13, January 2013.

[2] Hifaa Bait Baraka, Huaglory Tianfield,” Intrusion Detection System for Cloud Environment”, 7th International Conference on Security of Information and

Networks (SIN'14),IEEE 9-11 September 2014, Glasgow.

[3] Manish Kumar, Dr. M. Hanumanthappa," Scalable Intrusion Detection Systems Log Analysis using Cloud Computing Infrastructure", 978-1-4799-1597-2/13/ IEEE feb 2013

[4] H.Debar, M. Dacier, and A. Wespi, "Towards a taxonomy of Intrusion Detection Systems", The International Journal ofComputer and Telecommunications Networking -Special issue on computer network security, Volume 31 Issue 9, Pages 805 –822, April 1999,

[5] M. Roesch, Snort Lightweight Intrusion Detection for Networks, USENIX LISA, 1999.

[6] j. Gomez, C. Gil , N. Padilla , R. Banos, C. Jimenez, Design of a Snort-Based Hybrid Intrusion Detection System, *Proceedings of the 10th International Work-Conference on Artificial Neural Networks*, 2009,pp: 515-522.

[7] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler, "The Hadoop Distributed File System," IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), pp.1-10, 2010.

[8] T. White, Hadoop: the Definitive Guide, O'Reilly, 3rd ed., 2012

[9] Youngseok Lee; Wonchul Kang; Hyeongu Son, "An Internet traffic analysis method with MapReduce," Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP , vol., no., pp.357,361, 19-23 April 2010.

[10] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Cluster", OSDI'04 Proceedings of the 6thconference on Symposium on Opearting Systems Design & Implementation - Volume 6, Pages 10-10 ,USENIX,2004.

[11] Y. Lee, W. Kang, and Y. Lee, A Hadoop-based Packet Trace Processing Tool, International Workshop on Traffic Monitoring and Analysis (TMA 2011), April2011.

[12] http://www.sthurlow.com/python

[13] http://hadoop.apache.org/ (Accessed: 8 October 2013).

BIOGRAPHIES

Rashmi MR is an M.Tech scholar in School of Computer Sciences under M.G University of Kerala. She has Received B.Tech degree in Computer Engineering from SOE kalamassery under CUSAT University of Kerala. Her research interest includes Distributed Computing and Networking.

Dr. M Sudheep Elayidom is working as Associate Professor in the Department of Computer Science, SOE, CUSAT, and Kochi, India. He completed his PhD from CUSAT. His specialization is in Cloud Computing And Data Mining. He has published many research papers in National, International Conferences and Journals.

Dr. R VijayaKumar is working as Professor in SOCS, MG University. He completed his PhD from IIT Bombay. His specialization is in artificial intelligence and wireless communication. He has published many research papers in National, International Conferences and Journals.