iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** XII **Month of publication:** December 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Survey of Network Traffic Analysis Techniques

Hasmukh B. Domadiya[1], Dr. Girish C. Bhimani[2]
[1]*Assistant Professor, National Computer College, Jamnagar,*
[2]*Head of Department, Department of Statistics, Saurashtra University, Rajkot*

***Abstract:** A computer network is a set of computers connected with each other through a set of connections called a topology. Based on the area, a computer network could be a LAN, MAN or a WAN. The Internet is one of the widest computer networks composed of a several Wide Area Networks and Local Area Networks. Every user accesses the Internet through its LAN – Local Area Network. The usage of the Internet done by every user needs to be analyzed by the network administrator and ISP for various reasons discussed in this paper. With the growing world of digitalization and with the rapid increase of the Internet usage, an extremely large amount of data is generated to keep records of network traffic. Various techniques of analyzing network traffic are handled. Conventional methods are not suitable to process such extremely large volume of data. This paper shows how Hadoop based system can be used to perform network traffic analysis efficiently.*
***Keywords:** Internet, Network Traffic, Commodity Hardware, Hadoop, Map Reduce*

I. INTRODUCTION

A computer network is a collection of devices connected with each other for the purpose of communication. Most of the computer networks are layered based where each of the layers is responsible for facilitating communication in a specific way. Every communication needs to transfer some data from one computer to another. The Internet is also a computer network but it is huge and heterogeneous. The Internet is made of many small computer networks across the world. With the growing field of digitalization and rapid increase of Internet usage, a large amount of network traffic generated every second. It is necessary to analyze this data for accounting and auditing purpose. It is not possible to design a single centralized entity in the world which can process and analyze data of entire Internet. This is technically not possible due to extremely large data generated every second [1].

To handle such issue, several solutions have been proposed. The traffic data can be stored and analyzed at various levels for various purposes like accounting and auditing purpose. For example, an organization can build its own network with some special devices like firewall, log servers, authentication servers etc. Users inside an organization are under the supervision of these devices. Any user accessing the Internet from any of the devices connected with the organization network is subjected to be noticed and logged by these devices. User's all activities can be logged as well as allowed or disallowed using these devices. At another level, an ISP-Internet Service Provide may maintain certain devices to restrict access and keep logs of its individual customers. Such restrictions may be enforced based on policies of a country or of itself. This paper discusses why such network traffic needs to be analyzed. Network traffic analysis could be done at every layer – mostly at link layer, network layer, transport layer and application layer. This paper deals with performing network traffic analysis at or above application layer only. It also discusses some of the present solutions which are widely used [1].

II. NEED FOR TRAFFIC ANALYSIS

A computer network is a set of computers connected with each other through a set of connections called a topology. Based on the area, a computer network could be a LAN, MAN or a WAN. The Internet is one of the widest computer networks composed of a several Wide Area Networks and Local Area Networks. Every user accesses the Internet through its LAN – Local Area Network. The usage of the Internet done by every user needs to be analyzed by the network administrator and ISP for various reasons listed below [1].

A. Bandwidth Utilization

Bandwidth Utilization helps to find available speed to the Internet. Bandwidth utilization also helps in billing the users. Network administrator can restrict utilization according to the type of users. ISP can do billing or suggest plan according to the usage of any organization. ISP can also find top users who are consuming more bandwidth. Network administrator may set policies for limiting bandwidth utilization of users to avail required bandwidth for the working of critical applications.

Examples:

1) Decrease access speed once daily usage reaches to a limit.
2) Set dedicated bandwidth for the users of a video conference in a company.
3) Bill an organization based on its monthly usage.
4) Calculate penalty in case of unavailability of the Internet connectivity.
5) Detect usage of high bandwidth consuming applications.

B. Suspicious Activities

Every organization has a certain code of conduct related with the Internet access. Devices like firewall can be used to monitor and block illegal access. Network analysis helps in finding illegal attempts (pirated download, visiting blocked sites) done by various users. Such information helps a network administrator to prevent such users from accessing the Internet. Traffic analysis to detect suspicious activities related with some critical ports and protocols help to prevent various security related attacks too.

Examples:

1) Detect users who try to download pirated movies.
2) Block P2P applications to avoid torrent access.
3) Block TCP and UDP ports specific to ransom ware attacks.
4) Perform web and email filter to avoid access of malicious content.
5) Detect illegal attempts to access network resources like firewall, servers.

C. Usage Analysis

The web resources accessed by a user could be analyzed to find his area of interest. An organization could find recent trends by mining information searched by its users. Such information helps to find out what users most likely to do while using the Internet. Such information also helps to find out ranking of websites based on number of accesses.

Examples:-

1) Detect users who perform online shopping.
2) Find usage of local data servers across employees.
3) Analyze students' access patterns of an institute Moodle server.
4) Detect users who are busy with social networking most of the time.
5) Detect users who are using proxy sites to bypass security of network.

III. FIREWALLS

A firewall is a device which is a combination of hardware and software. The purpose of firewall is to control inbound and outbound traffic for a local network. For example, a company can have its network using internetworking devices like routers, switches, access points etc. The company has purchased an Internet link which is connected with the firewall. The rest of the network is also connected with the firewall. So here a firewall works as a connecting point between the Internet and the local network. Any traffic needs to be passed and permitted by firewall. One such scenario is shown in Figure 1 [1].

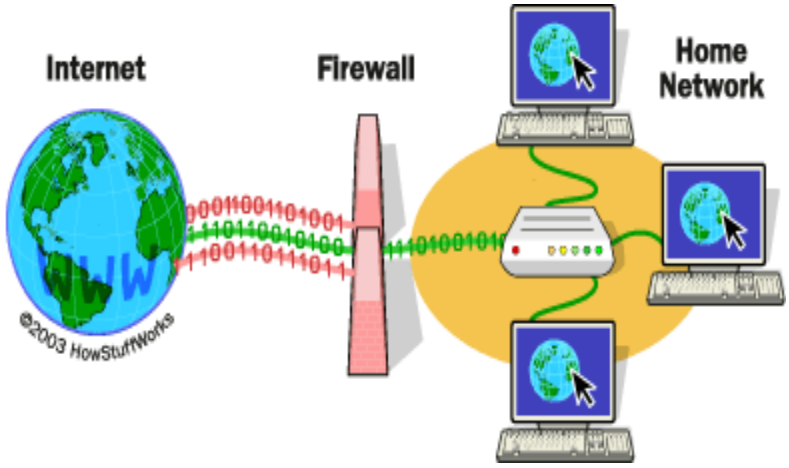


Figure – 1 Firewall

Here we can see that the home network and the Internet are connected through firewall. The main purposes of having firewall are listed below [1].

A. User Management

Firewall facilitates user authentication to access the Internet and other LAN resources like local web servers, directory servers etc. The main purpose of authentication is to manage logs based on user login. Firewalls support various types of users like local users, remote users, LDAP users etc. Credential database for local users is stored in the firewall itself. Credential database for remote users is stored at an authentication server like AAA servers. LDAP- Lightweight Directory Access Protocol based authentication uses a LDAP server to manage user credentials. Other than authentication, the purpose of user management is to set policies to provide access rights. For example, in an organization, among the employees, network admin sets some policies like Managers have full access of the Internet (including social networking websites and applications) while workers have no access of social networking websites and applications). Similarly policies can be set to set bandwidth policies (to restrict speed at which a user can access the Internet), usage policies (to restrict amount of data a user can access), time policies(to set schedule for user access) etc.

B. Log Management

Firewall facilitates recording logs of various activities done by various users. Most of the firewall support various way of keeping logs such as by user, by IP address or by MAC address too. The purpose is to provide auditing facility where a network administrator can look into the logs to determine who had accessed what Internet resources at what time and for how long. Every firewall manages logs into its own hard disk while in a corporate intranet; a separate log server can be used to store and maintain log records beyond the capacity of a firewall. Logs are automatically maintained and deleted based on first come first out basis. Many firewall vendors provide cloud storage to automatic integration and submission of logs to the cloud for storing logs on long term basis. The format of log is firewall dependent but some fields are common like user identity (username, IP address, MAC address (If firewall supports)), Date and Time, URL etc.

C. Filtering

As the name suggest, firewall is meant to be a protective device which prevents unauthorized traffic flow based on the policies set by network administrator. Filtering could be done at various levels like user filtering (disallows blocked or banned users to access the Internet), web filtering blocks access to prohibited websites (websites of social networking, pornography, pirated content) are blocked. Antivirus blocks any traffic which is having viruses etc. The filtering system also makes logs so that later on we can find out which user had tried to access unauthorized and prohibited content.

IV. ISPs

An Internet Service Provider provides Internet connectivity to individual customers. A customer can be a home user or corporate users – owning his own intranet. ISP has to keep logs of user usages for mainly two purposes which are explained below.

A. Accounting

Most of the ISP charge their customers based on their usage. Various ISPs provide various schemes based on the speed and amount of data customer use. For example, ISP keeps logs of how many GBs a post paid customer has used so far. Based on this information, customer is provided a bill for it. For a customer who has opted for an unlimited plan where after reaching a specific limit, its speed is reduced, ISP keeps records of daily usage and limits speed after reaching the daily limit. Some ISPs provide services where the down time is measured and concession is given to the customer based on unavailability of services after a specific amount of time.

B. Auditing

As discussed earlier in section III, the scope of a firewall is to serve only local users inside a corporate intranet. But in case of ISP, it has to serve different customers of different type (home user or a corporate user owning his own network). Most of the normal internet plans, ISP provides dynamic IP address based services where a dynamic IP is assigned to a customer once he tries to start accessing the Internet. In a few special – corporate plans, ISP provides static IP addresses too. On the whole, ISP maintains which IP address was provided to which customer and what Internet a customer has accessed. A nature question arise is whether ISP can get information about encrypted content or not. the answer is ISP is always able to store hostname (URL) or the IP address (of the remote computer) irreverent to the encryption status of the content. Here auditing is at the next level, ISP does not see which local user has accessed which resource because local users are visible up to the firewall only. At the ISP, every access is done by a IP

address assigned to the corporate intranet only. ISPs always analyses customer activities and take necessary actions if a customer violates policies directed by the Government.

V. SNIFFERS

Sniffers are some software tools which can be used to analyze packets travelling in a network. These tools can be easily installed and used to monitor what's going on in a network. Nowadays some sniffers are combination of hardware and software which provide facilities to access real time network packets as well as storing their statistics on hard disk for future usage. There is always a debate over purpose of using sniffers. Earlier in the days of networking, sniffers were used by only network qualified people. Nowadays sniffers are easy to use and freely available. Hackers use them to obtain passwords of systems where encryptions are not made. A curious user uses sniffers to get into the detail of what's going on. Network administrators use them to test filtering and other security features. Wireshark, tcpdump, commonview, capsa, networminer are some of the most popular tools. In older days, sniffers were able to capture packets only intended for the computer where they are running. In recent sniffers, promiscuous mode makes all traffic in a network available to the computer even when not directly indented for it [2,3,4]. The usual location of a sniffer in a network is shown in Figure 2

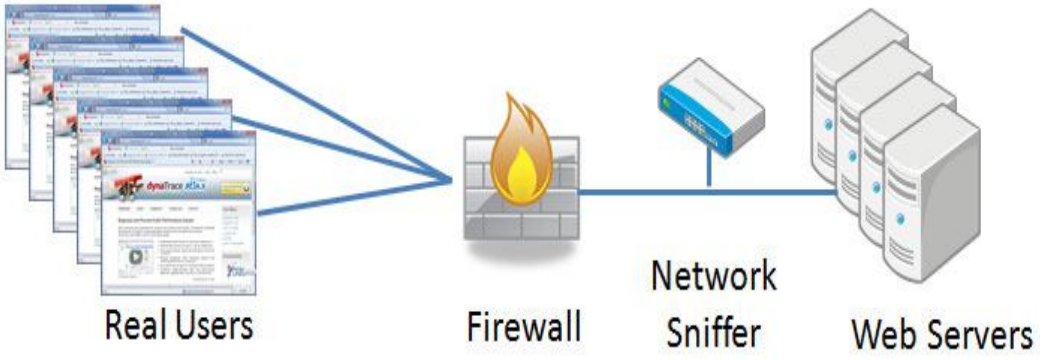


Figure 2 – Sniffer

VI. HADOOP

A. *Disadvantage of Centralized Systems*

Every network keeps logs of all the usage made by various users through devices like firewall. Data is kept at various layer of network architecture too. In the era of computerization, the number of users of every network increases rapidly. Ultimately the number of users of the Internet is increasing rapidly. Computerization increases web resources too. In such environment, every network has to maintain a large data to store logs related with the network traffic. Such data keep increasing with time leading to consuming storage in GBs and even in TBs. It is difficult to analyze such huge amount of data with the help of conventional file system and centralized processing [5.6].

Every network has a hierarchy of network devices like firewall, layered switches, servers etc. lets discuss what happens to log network traffic in a traditional way. Firewall has its own storage where it stores logs. Its internal or external analyzer can read the logs and provide analytics to the network administrator. In another scenario, data servers could be used to keep logs stored. Traditional database management systems could be used to design application programs to analyze these logs. Both the ways are suitable if the amount of traffic is low to moderate. With special storage based products, this approach is suitable for high amount of traffic too. This way of handling network traffic fails when working with extremely high amount of traffic. In reality, large scale networks generate logs at extremely increasing rate leading to logs of GBs and TBs. traditional way of processing data fails due to its own limitations. Here are some of the disadvantages while using traditional way of processing for extremely high amount of traffic (We could call it as Big Data) [7,8].

1) Centralized storage needs a large amount of space. Redundancy through replications is more costly too.
2) Centralized processing needs extremely powerful processing units. Preferably a large scale parallel computer can be used.
3) Not suitable if data itself is distributed due to heavy overload during data transfer.
4) Programs with extreme parallelism are required to provide timely analytics.

B. Hadoop as a Solution

Hadoop supports storage and processing of extremely large amount of data (BigData) in distributed computing environment. Data could be spitted at various physically and geographically separated nodes called data nodes. These parts are configured as a single file system called HDFS (Hadoop Distributed File System). A name node could manage task assignment and retrieval of analytics over it. Hadoop suits well in two situations. These two situations are true while analyzing network traffic [9].

1) Huge amount of data to process. (GBs, TBs)

2) Data needs to be processed with the concept of write once, read many times requirements.

Doug Cutting, Mike Cafarella and started an Open Source Project called Hadoop based on using Map Reduce. Now Apache Hadoop is a registered trademark of Apache Software Foundation. Hadoop runs applications using the MR-Map Reduce programs. Hadoop uses commodity hardware based system to process data in parallel and distributed manner. Hadoop is capable to design applications to process huge amount of data for statistical analysis purpose [10]. A Hadoop based system is shown in Figure 3.

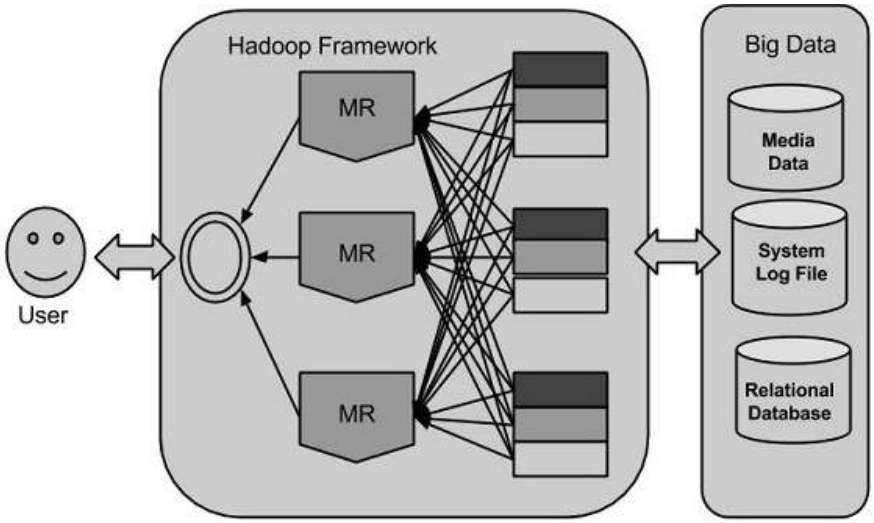


Figure 3 Hadoop

VII. CONCLUSION

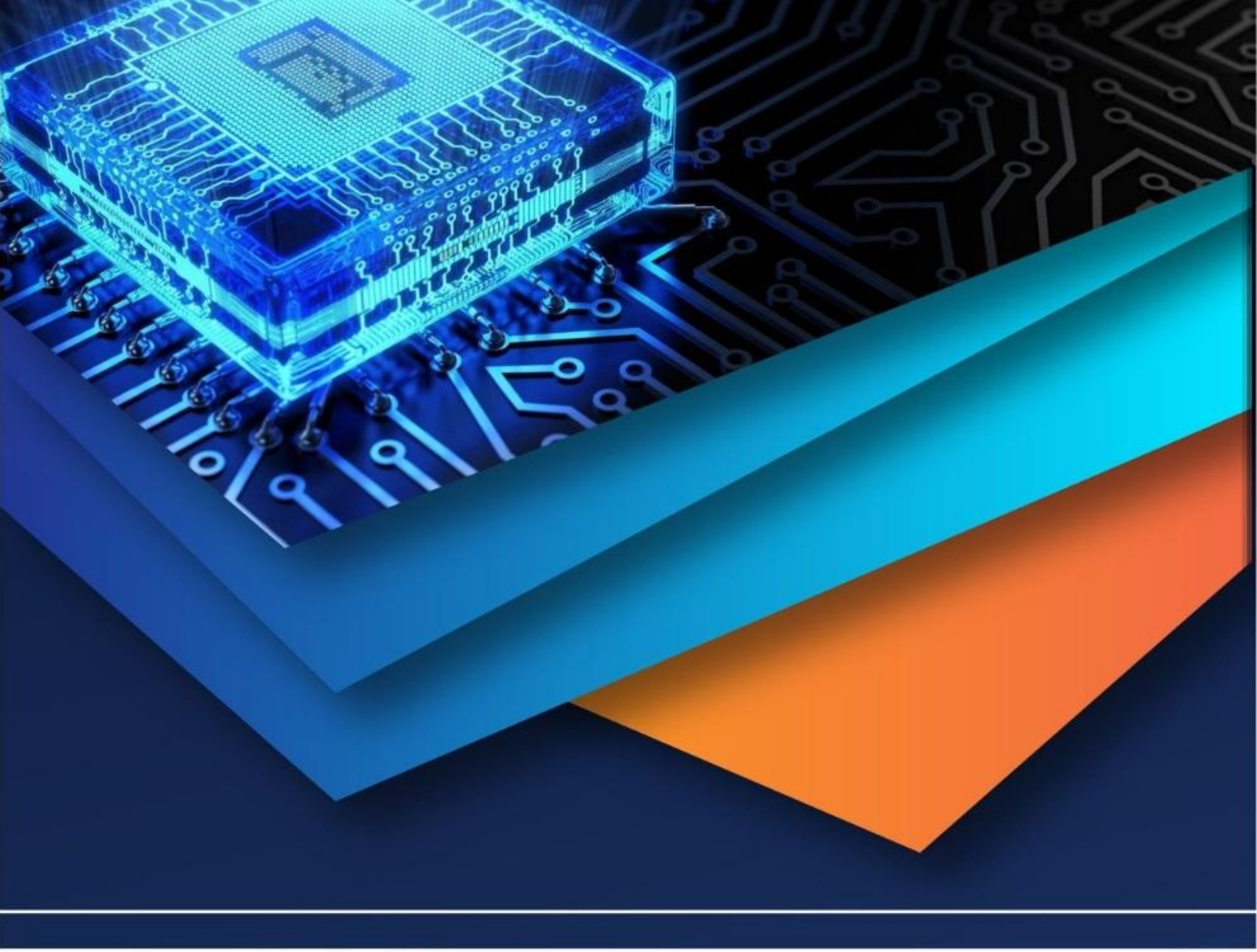
This work discusses why and how every network – which is connected with the Internet, needs to be analysed as far as its traffic is concerned. Various purposes for network traffic analysis are explained. Various traditional approaches to perform such analysis are explained which are firewall based, ISP based and sniffer based. As discussed firewall based network traffic is completely centralized and limited to do analyze within a corporate intranet only. ISP based solution refers to the public IP only and it does not look into the local users. It considers a corporate intranet as a single logical user and analysis is not performed on which user of intranet has accessed what resources. Sniffers are powerful but have limited analysis facilities and they are suitable to do analysis within a network. With the growing usage of computerization, it is one of the obvious needs to process large volume of data efficiently. Hadoop is the solution to process large volume of data efficiently. Rapidly increase usage of the Internet invite a challenge to perform traffic analysis on large scale basis – amount of traffic is extremely increasing day by day which increases amount of logs day by day. It is been shown that how Hadoop based system could be useful to analyse large amount of network traffic.

VIII. FUTURE WORK

Further research work can be carried out towards developing Hadoop based solutions to mine network traffic to find out useful information. If a corporate sector has multiple branches across multiple cities and each of the branches is having its own intranet controlled by a firewall, it is possible to perform network traffic analysis among all the logs maintained at all of these firewalls using Hadoop. Hadoop enables local processing at every intranet to reduce transfer of high volume of log data to a centralized system. This is possible using formulation of a Hadoop cluster. Even at a standalone processing with a centralized system Hadoop based Map Reduce programming are faster as far as processing large volume of data is concerned. Map Reduce based java programs are comparatively easy to write and efficient as compared to traditional programs.

REFERENCE

[1] Law K. SE 4C03 Winter 2005 An Introduction of Firewall Architectures and Functions. 2005.
[2] Tcpdump, http://www.tcpdump.org.
[3] Wireshark, http://www.wireshark.org.
[4] CAIDA CoralReef Software Suite,http://www.caida.org/tools/measurement/coralreef.
[5] Hadoop, http://hadoop.apache.org/.
[6] White, Tom. Hadoop: The definitive guide. " O'Reilly Media, Inc.", 2012.
[7] Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010
[8] J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Cluster, OSDI, 2004.
[9] Lee, Yeonhee, and Youngseok Lee. "Toward scalable internet traffic measurement and analysis with hadoop." ACM SIGCOMM Computer Communication Review 43.1 (2013): 5-13.
[10] Lee, Youngseok, Wonchul Kang, and Hyeongu Son. "An internet traffic analysis method with mapreduce." Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP. IEEE, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY