Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 17 Issue 2 Version 1.0 Year 2017
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Measurement and Classification of Network Traffic Analysis using Hardoop

By Tanmay Paul
Adamas Institute of Technology

Abstract- Network traffic can classified as a process which list computer network based on some parameters like port number and protocols into some traffic classes like undesired, sensitivity etc. Traffic can be implemented differently to differentiate the service required for the user for the specific purpose. The large demand of increase in internet users and increase in bandwidth required for various applications are escalating day by day. The traffic data needs to be classified and analyzed with certain tools. Hardoop is the tool which performs the task in very time efficient manner. Hardoop actually run on commodity hardware which process this huge data with hive. Traffic analysis, measurement and classification are done by hardoop based tools at various parameters of packet and flow level. The derived result is used by network administrator for resolving networking related issue. The measurement of internet traffic and analysis has been implemented from long before but the problem is recent years the user in internet has escalated dramatically. We proposed network traffic management system for analyzing internet traffic of multi-terabytes in extensible manner to perform HTTP, ICMP, UDP, TCP and IP.

Keywords: *network traffic, hardoop, traffic management and analysis, HDFS, HIVE, IP.*

GJCST-C Classification: *C.2.3*

Strictly as per the compliance and regulations of:

Measurement and Classification of Network Traffic Analysis using Hardoop

Tanmay Paul

Abstract- Network traffic can classified as a process which list computer network based on some parameters like port number and protocols into some traffic classes like undesired, sensitivity etc. Traffic can be implemented differently to differentiate the service required for the user for the specific purpose. The large demand of increase in internet users and increase in bandwidth required for various applications are escalating day by day. The traffic data needs to be classified and analyzed with certain tools. Hardoop is the tool which performs the task in very time efficient manner. Hardoop actually run on commodity hardware which process this huge data with hive. Traffic analysis, measurement and classification are done by hardoop based tools at various parameters of packet and flow level. The derived result is used by network administrator for resolving networking related issue. The measurement of internet traffic and analysis has been implemented from long before but the problem is recent years the user in internet has escalated dramatically. We proposed network traffic management system for analyzing internet traffic of multi-terabytes in extensible manner to perform HTTP, ICMP, UDP, TCP and IP.

Keywords: network traffic, hardoop, traffic management and analysis, HDFS, HIVE, IP.

I. Introduction

The collection of different servers, computers, peripherals, devices when connected to one another for secure mean of communication is described as network which is mainly used for sharing data, or as a means of communication. The process of monitoring network traffic involves managing and analyzing network to overcome any discrepancy that might be a problem for the network. The amount of data involve in communication between network is described as network traffic. The network packet [1] mostly comprised of network data which makes the load within the network. The monitoring mainly involves analyses incoming and outgoing packet. The measurement of traffic over a particular network is called traffic measurement. There are basically two types of techniques involved. Firstly the active techniques and the secondly is the passive technique. Active [2]are more accurate and instructive and the main drawback is that it may over crowd the network by infuse with artificial inquest traffic whereas passive [3] run on the background which can be used to implement network analyzing action and the drawback is that supervise on all network[4]. The main challenge of internet traffic measurement is firstly flow statistics computation time and secondly single node failure. To overcome this problem we implement [5] hardoop framework. Hardoop is actually an open source software framework for large set data processing and storage. It provides necessary possibilities of scaling and fault tolerance which are the most important in networking. We also implement Map Reduce model to resolve the inconsistency in between the hardoop data distribution and network monitoring where data is recorded and splinted and dispense them into cluster for individual processing. The related packets may spread across different splits, thus dislocating traffic structures that are essential for network traffic monitoring. In this paper we have proposed a novel method for network traffic measurement and analysis.

Author: Department of Computer Science, Adamas Institute of Technology, Kolkata, India. e-mail: tanmaypaulcs@gmail.com

II. Software Overview

In Hardoop we can analyses and process large data set. It eliminated the use of expensive hardware for storing and analyzing huge data. It minimizes the cost of installing distributed parallel processing of the data by installing hardware in existing servers. By implementation of hardoop it enables to process and analyses the data more efficiently and also by reducing the cost. It also enables the organization to import and use the data one became absolute. Below in Fig 1 the flow chart of data flow of 7 layers of OSI model of traffic analysis based on hardoop is given. The tools mentioned below in Table 1 are efficient of data analyzing but are limited to storage and measurement. The traffic sampling method can be used to overcome the limitation where results are drawn through partial observation. The implementation of SQL is also not proposed due to its nature of query operation. Below in Table 1 networking traffic monitoring tools are given with uses and limitations are described.

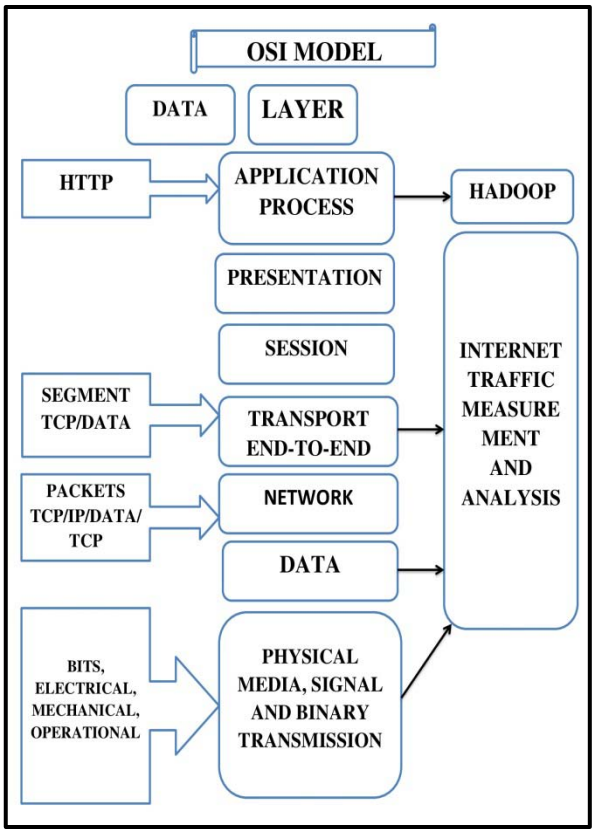

Fig.1: OSI model based on hardoop (source: 6)

Table 1: Networking traffic monitoring tools.

TOOL NAME	OPERATING SYSTEM	LANGUAGE	USE	DISADVANTAGE
NETWORK MINER	Windows, Mac, Linux, FreeBSD.	C	Used as passive network sniffer/packet capturing tool in order to detect OS sessions, host name, open ports etc.	Cost is high about 70$
WIRESHARK	Linux, OS X, BSD, Solaris, windows.	C, C++	It allows examining of the data from a live network or from a capture file on disk. The data can also be browsed and delving down into packet level as required	The main issue is the security features of this tool.
TEPDUMP	Unix like OS, Linux, OS X, BSD, Solaris, windows, android and AIX	C	The user with the necessary privilege acting on a router or gateway through which unencrypted traffic such as telnet passes can use Tepdump to view login id, passwords, URLs, content of website being viewed or any other unencrypted information	It does not receives new features update and keep resolving the bugs and troubleshooting the previous networking issues.

III. System Overview

The system proposed involved firstly input conversion, secondly hardoop pre-processing and qlikView [7] analysis. At first for the packet capture jpcap and wincap [8-9] is used for capturing which is used for supporting the jdk environment and wincap supports the window environment. After capturing the packet gets converted into .text file or .csv file for training data. The dataset made gets loaded as input for category. The processed file is stored in HDFS and to represent in HIVE file externally. And at last IP analysis, port no, protocol and displayed in graphical format. Below in Fig2 the work flow diagram of the proposed system has been given.

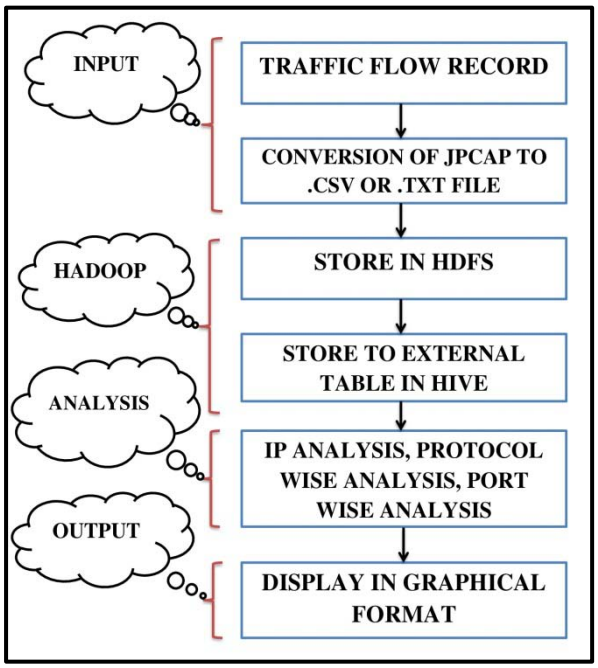

Fig.2: Work flow diagram

Hardoop has been implemented for network traffic analysis and measurement. The various characteristic of traffic data is been considered as IP address for traffic counting, total traffic data size, traffic counting with port based classification where total traffic and size per port is calculated. The internet traffic is being captured from Adamas institute, Kolkata which has been stored in jcap and wincap format. The slave node stores the data with the replication factor of 3 which means 1 file is stored and min Fig3 the network diagram has been given.

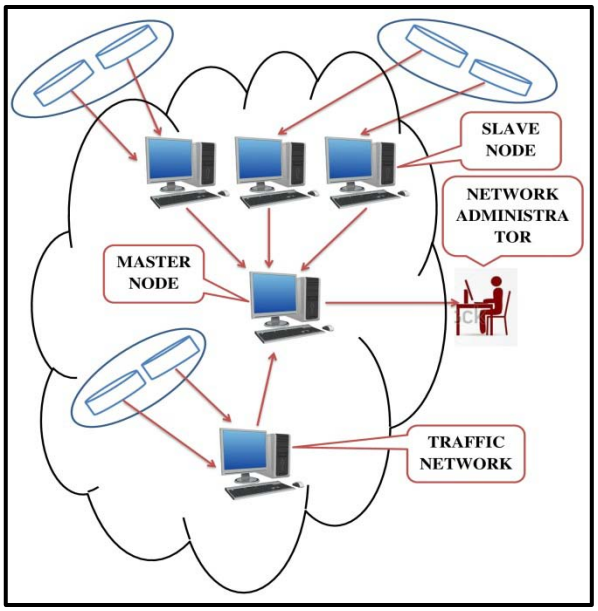

Fig.3: Network diagram.

IV. Experimental Evaluation

Protocol based network packet are captured, port number having LAN making use of java API.2 and IP addresses. The captured file stored in HDFS [10-11] is described data wise. The top 10 IP address can be calculated to define the user usage so that the network which consumes more traffic or more bandwidth can be identified. The total number of packet has also been calculated based on port which his called port-wise byte counts. Port 443 (HTTPS) having the highest number of count which is about 59% has also been shown below. The size of packet and total number of packet each day has been calculated. Below in Fig4 the top 10 IP address usage is shown and in Fig5 the port wise byte count is also shown.

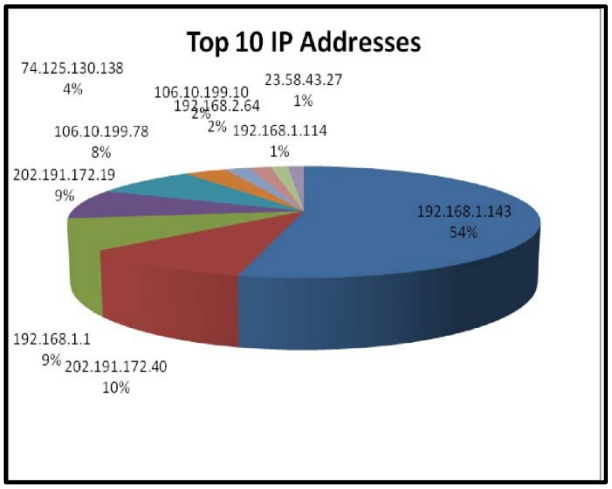

Fig.4: The top 10 IP address data usage.

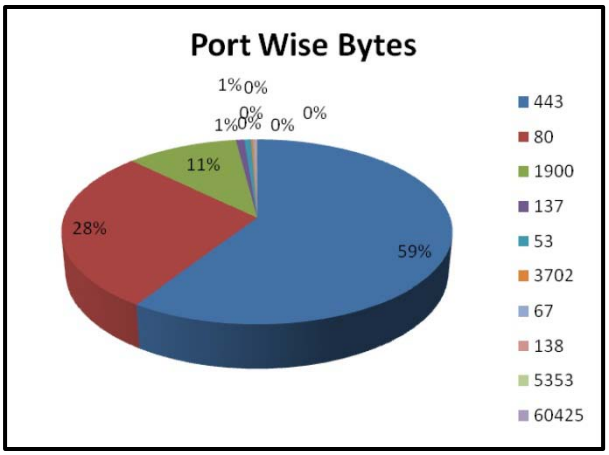

Fig.5: The port-wise byte count.

V. Conclusion

The network traffic analysis we proposed in this paper will be very efficient for the network administrator to monitor the bandwidth consumption and maintain the system and trouble shoot bugs if necessary. In the paper our main focus was on the flow packet and analysis by network topology. The huge amount of data cannot be handled with single server so large dataset is necessary for matching the computing and storage, and scalable analysis becomes a problem. That the reason we introduce Hardoop as an open-source platform which resolves all the issue in large data set analysis. We have proposed the novel method of data analysis and measurement.

References Références Referencias

1. Wu Kehe; Cheng Rui; Zhang Yingqiang; Mu Hongtao "The research on the software architecture of network packet processing based on the many-core processors":2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS).
2. Ali R. Ahmadi; Laura Kane; Robert Macdonald; GrahamAult; Pedro Almeida; Sotiris Georgiopoulos; JoseBarros; Panagiotis Papadopoulos "Active network management supporting energy storage integration into system, market and the distribution network":CIRED Workshop 2016.
3. Shuonan Shang; YongqingMeng; Jian Wang; Huixuan Li; Wei Ren; Xifan Wang; Yong Cui "Research on modeling and control strategy of modular multilevel matrix converter supplying passive networks" : 2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC).
4. Liu Mingbo; Sun Wenjie; Zhao Qianhong; Tian Zhaoping "Design and implementation of IP network traffic monitoring system": 2016 15th International Conference on Optical Communications and Networks (ICOCN).

5. Vaibhav Fanibhare; Vijay Dahake “Smart Grids Map Reduce framework using Hadoop”: 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN).
6. Lena T. Ibrahim; Rosilah Hassan; Kamsuriah Ahmad; AsrulNizamAsat “A study on improvement of internet traffic measurement and analysis usingH adoop system”: 2015 International Conference on Electrical Engineering and Informatics (ICEEI).
7. Roumianallieva; KirilAnguelov; Delyana Gashurova “Monitoring and optimization of e-Services in IT Service Desk Systems”: 2016 19th International Symposium on Electrical Apparatus and Technologies (SIELA).
8. Yong Xing Wang; Xiu Zhu Jiang; Chun Wang “Design and simulation on the PC of routing software based on Wincap”: The 2011 IEEE/ICME International Conference on Complex Medical Engineering.
9. Wenjian Xing; Yunlan Zhao; Tonglei Li “Research on the Defense Against ARP Spoofing Attacks Based on Winpcap”: 2010 Second International Workshop on Education Technology and Computer Science.
10. Mao Ye; Jun Wang; Jiangling Yin; Xuhong Zhang “Accelerating I/O Performance of SVM on HDFS”: 2016 IEEE International Conference on Cluster Computing (CLUSTER).
11. Bikash Agrawal; Raymond Hansen; Chunming Rong; Tomasz Wiktorski “SD-HDFS: Secure Deletion in Hadoop Distributed File System”: 2016 IEEE International Congress on Big Data (Big Data Congress).