IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Network Traffic Analysis Measurement and Classification Using Hadoop

Preeti Joshi[1], Arati Bhandari[2], Kalyani Jamunkar[3], Kanchan Warghade[4], Priyanka Lokhande[5]

Assistant Professor, Information Technology, MMCOE, Pune, India [1]

Student, Information Technology, MMCOE, Pune, India [2,3,4,5]

Abstract: Traffic classification is a process which categorizes computer network traffic according to various parameters (for example, based on port number or protocol) into a number of traffic classes such as Sensitive, Best-Effort, and Undesired etc. Each resulting traffic class can be treated differently in order to differentiate the service implied for the user. Due to growth in Internet users and bandwidth-hungry applications; the amount of Internet traffic data generated is so huge. It requires scalable tools to analyze, measure, and classify this traffic data. Traditional tools fail to do this task due to their limited computational capacity and storage capacity. Hadoop is a distributed framework which performs this task in very efficient manner. Hadoop mainly runs on commodity hardware with distributed storage and process this huge amount of traffic data with a Hive. Hadoop-based Traffic Analysis Measurement and Classification tool which perform Traffic Analysis, Measurement, and Classification with respect to various parameters at packet and flow level. The results can be used by Network Administrator and ISP's for various usages. Internet traffic measurement and analysis has long been used to characterize network usage and user behaviours, but faces the problem of scalability under the explosive growth of Internet traffic and high-speed access. We proposed a traffic monitoring system that performs IP, ICMP,TCP, HTTP, and UDP analysis of multi-terabytes of Internet traffic in a scalable manner. This can achieve the performance challenges such as accuracy, scalability.

Keywords: Internet Traffic, traffic measurement and analysis, HDFS, HIVE, Qlikview or Tableau.

I. INTRODUCTION

A network is a collection of computers, servers, mainframes, network devices, peripherals, or other devices connected to one another. A network is used for sharing resources, exchange files, or allow electronic communications. Network traffic monitoring is the process of reviewing, analyzing and managing network traffic for any irregularity or process that can affect network performance, availability or security. Network traffic is nothing but the amount of data moving across a network at a given point of time. Network data is mostly enclosed in network packets, which forms the load in the network. Network traffic is the main component for network traffic control, network traffic measurement and simulation. The main objective of network traffic monitoring is to ensure availability and smooth operations on a computer network. Network monitoring includes network sniffing and packet capturing techniques in monitoring a network. Network traffic monitoring mostly requires reviewing each incoming and outgoing packet. The amount of Internet traffic data generated is ever increasing, due to the growth in Internet users and bandwidth-hungry applications. It requires scalable tools to analyze, measure, and classify large amount traffic data. A network traffic can be classified using HTTP, IP,ICMP,TCP and UDP analysis of multi-terabytes of Internet traffic[1].

In computer networks, network traffic measurement is the process of measuring the amount and type of traffic on a particular network. Network analysis could be measured by active technique and passive techniques. Active techniques are more intrusive but are arguably more accurate. Passive techniques are of less network overhead and hence can run in the background to be used to trigger network management actions. A limitation of active measurement is that it may disturb the network by injecting artificial probe traffic into the network and the main drawback of using this passive measurement is that it assumed that it "owns" all networks [4].

In the network traffic measurement there are mainly two challenges like:
1) Flow statistics computation time and
2) Single node failure.

To address these challenges, we implemented the internet traffic measurement and analysis using Hadoop framework. Apache Hadoop is an open source software framework for storage and large scale processing of datasets.

Motivated by the need for analyzing large datasets, we investigate the benefits of hadoop to achieve real time network traffic. However, available implementation such as Hadoop provides scaling capabilities and fault tolerances that are essential features for Internet security.
However, we found fundamental inconsistency between Hadoop data distribution and network traffic monitoring. In the MapReduce model, data is conceptually record-oriented; sequentially, Hadoop divides datasets into splits i.e., subsets of records, and distributes them in a cluster to be independently processed. Splits of a dataset have all the same size to make sure that the processing end time for all nodes coincides. In the case of network traffic, a split usually represents a subset of packets. However, related packets may spread across different splits, thus dislocating

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

traffic structures that are essential for network traffic monitoring. For example, network traffic monitoring tools can identify HTTP protocols because of the numerous requests sent from application layer.

The organization of this paper is as follows: Chapter 2 focuses on the overview of hadoop. Chapter 3 focuses on related work done in traffic analysis with its pros and cons. Chapter 4 gives a detailed description of system overview. Chapter 5 gives detailed description of proposed work with implementation. Chapter 6 gives experimental setup and results of proposed system. At last Chapter 7 concludes the paper and focuses on future direction to our work.

II. OVERVIEW OF HADOOP

Hadoop is an open source platform established to collect, manage and process a large dataset. Traditionally, organization heavily depends on expensive, proprietary hardware and huge storage systems to store and process data.

However, Hadoop minimizes the needs of organizations to invest heavily on replacing or expanding the hardware by enabling distributed parallel processing of huge amounts of data across existing servers. By implementing Hadoop, it enables the organization to harness the ability to store and process the data faster and efficiently by adding new hardware without limits. This is an advantage as there is a need for organizations to collect data from new sources from day to day. The ability to invoke organizations to keep and find value to data which was once considered worthless [3].

The Figure: 1 explains the overview of data flow through 7 layer OSI model based upon the Hadoop internet traffic analysis.

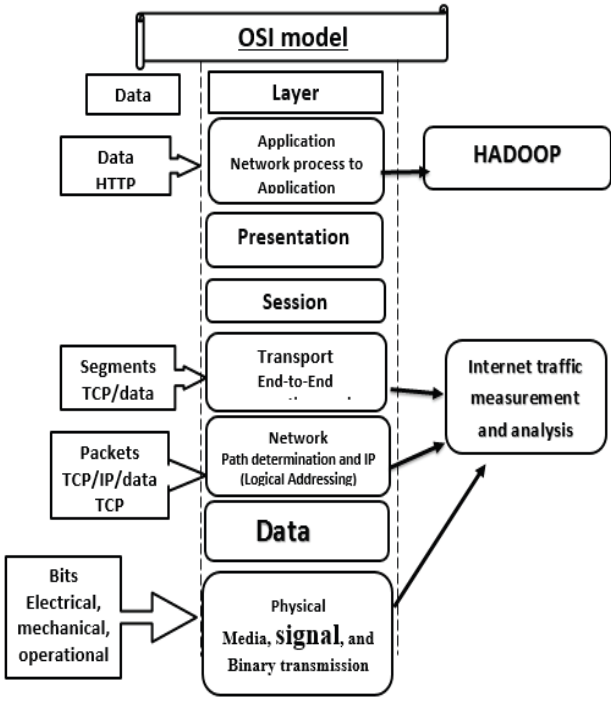

Figure1: Overview of the model [9]

III. RELATED WORK

This chapter gives the idea of how different network traffic tools are working on their environment. Also we classify existing approaches for Internet traffic measurement and analysis with their advantages and limitation.

In Internet traffic measurement and analysis, flow-based traffic monitoring methods are widely deployed throughout Internet Service Providers (ISPs), because the volume of processed data is reduced and many convenient flow statistics tools are available.

The tools from Table 1 are good for analysis but restricted to storage capacity and processing power capability. To overcome these restrictions, traffic sampling can be used where partial observations are made to draw results but it will result in loss of information. The traditional relational database using SQL is also impractical due to sequential nature of query operations [4].

If we distribute the traffic data among multiple nodes there are chances of failure of certain machines in distributed environment hence, data availability becomes a critical issue also in traditional tools the fault-tolerance issue is not handled. Therefore it is necessary to develop tool which overcomes all the problems faced by older tools [12].

Tool name	Operating system	Language used	Latest release	Use	Disadvantage
Network Miner	Windows (but also works in Linux / Mac OS X / FreeBSD)	C#	version 1.0 on Feb. 5, 2011	Used as a passive network sniffer/packet capturing tool in order to detect operating systems, sessions, hostnames, open ports etc.	Cost is so high i.e. $70
Wireshark	Linux, OS X, BSD, Solaris, some other Unix-like operating systems, and Microsoft Windows.	C,C++	version 1.12.7 on Aug. 12, 2015	It allows you to examine data from a live network or from a capture file on disk. You can interactively browse the capture data, delving down into just the level of packet detail you need.	Wireshark has suffered from dozens of remotely exploitable security holes, so stay up-to-date and be wary of running it on untrusted or hostile networks (such as security conferences).
Tcpdump	Unix-like operating systems: Linux, Solaris, BSD, OS X, HP-UX, Android and AIX	C	version 4.7.4 on April 22, 2015	A user with the necessary privileges on a system acting as a router or gateway through which unencrypted traffic such as Telnet or HTTP passes can use tcpdump to view login IDs, passwords, the URLs and content of websites being viewed, or any other unencrypted information.	It does the job well and with fewer security risks. It also requires fewer system resources. While Tcpdump doesn't receive new features often, it is actively maintained to fix bugs and portability problems
inSSIDer	Windows, OS X, and Android	C	version 4.1.0 on Jan. 22, 2015	inSSIDer can find open wireless access points, track signal strength over time, and save logs with GPS records	Not working well on 64-bit Windows and Windows Vista.

Table 1: Network traffic monitoring tools

A. Big Data Analytic

Organizations using Big Data analytic software are deploying it on massive parallel clusters available in the market. Due to cost consideration, organizations have opted on leveraging an open source framework, such as the Apache Hadoop to distributed file systems.

B. Internet Traffic Measurement and Analysis

Most of Internet traffic measurement and analysis tools are unable to handle a huge traffic and failed at high-speed links of routers in scalable means.

The Apache Hadoop software library uses a straightforward model of programming which allows the distributed processing of extensive amount data sets

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

throughout information processing system's mass. The library is capable to restore and manage failure at the stage of application instead of depending on hardware to convey peak-availability [15]. Hadoop provides computing and storage mechanisms of fault-tolerant for huge-scale cluster surroundings.

IV. SYSTEM OVERVIEW

Module Description:
The proposed work involves three main phases: Conversion of Input, Pre-processing by Hadoop and Analysis using QlikView or Tableau. All these phases which constitute the working procedure of the system is represented by the following flow chart in Fig. 2.

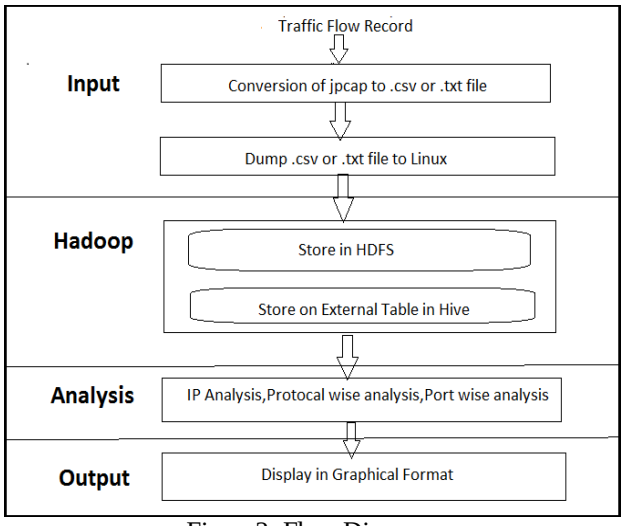

Figure2: Flow Diagram

A. First Phase
i] Packet Capture:-
We are using *jpcap* and *winpcap* to capture packets. *jpcap* is used for supporting jdk environment. *winpcap* is used for supporting to windows environment.
ii] Training for live packets:-
After capturing live packets they are converted into .text file or .csv file, to be used as training data.
iii] Import Dataset:-
This newly created training dataset is loaded into system as input for classification.

B. Second Phase
The second phase stores and processes the .csv or .txt file in HDFS and data is represent in external Hive table.

C. Third Phase
The third phase includes analysis of IP address, protocol, port no. and output is display in graphical format by using Tableau or QlikView.

V. PROPOSED SYSTEM

By considering all the problems of older tools, we have proposed Network traffic analysis measurement and classification using Hadoop. It considers various aspects of traffic data such as IP address wise traffic count, total Size of traffic data, date-wise traffic count along with port based classification where total traffic and total size per port is calculated.

We are going to capture Internet traffic from router of our MMCOE Campus, Karvenagar which is basically stored in jpcap or winpcap format as per specification. The Slave Nodes i.e. Data Nodes stores this traffic data with replication factor of 3 means one file get stored onto 3 different slaves for data availability.

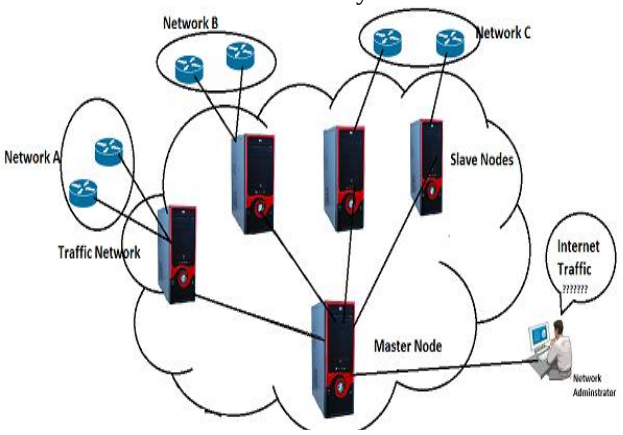

Figure3: Block Diagram

The proposed algorithm is as follows:

Algorithm: Network Analysis

Input: Traffic captured in jpcap or winpcap format.
Output: txt file.
Begin
Step 1: Method: start capture.
Step 2: Convert jpcap file format into txt file format.
Step 3: Store traffic data onto HDFS.
Step 4: Analysis done by HIVE Query.
Step 5: Calculate
a. source_ip_wise_traffic_count
b. destination_ip_wise_traffic_count
c. port based classification
d. Date_wise_traffic_count
e. protocol_wise_traffic_count
Step 6: Display result into graphical format.
Step 7: End

VI. EXPERIMENTAL RESULTS AND ANALYSIS

1. Network packets are captured based on protocols, ip addresses, port no from LAN using Java API.

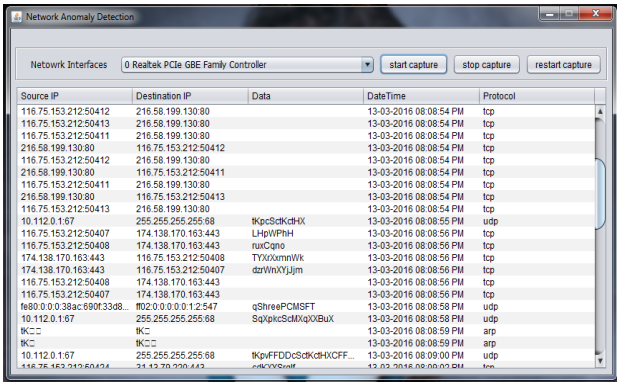

Figure 4: Packet Capturing

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

2. File stored in HDFS: Captured Packets are stored in HDFS which is classified date wise.

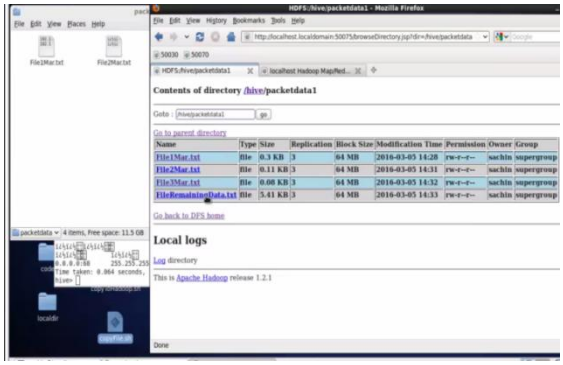

Figure 5: File Storage in HDFS

3. Top 10 IP Addresses: We can calculate the top 10 users from the packets captured who generates more traffic as shown in figure 6. This information can be used for identifying users which are consuming more bandwidth.

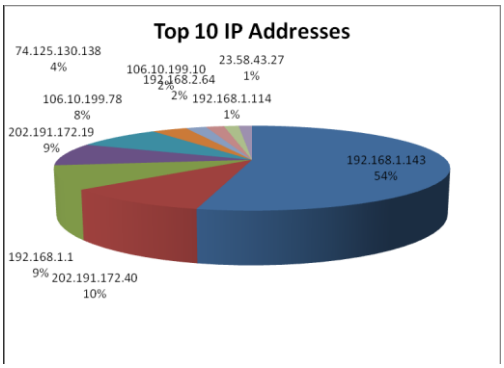

Figure 6: Top 10 IP Addresses

4. Port-Wise Byte Count:
We have also calculated the total number of packets port number wise . Port 443 (HTTPS) having higher number of byte count. The results are shown in figure 7. We have also calculated number of packets per day and size of packets per day.

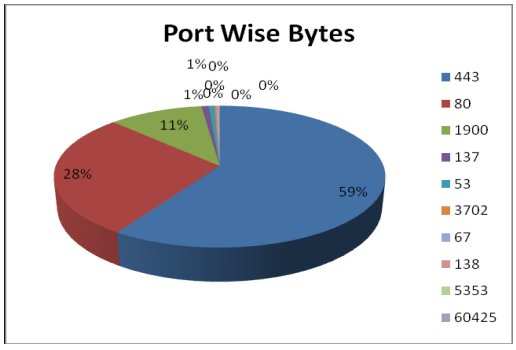

Figure 7: Port-Wise Byte Count

VII. CONCLUSION AND FUTURE WORK

The network traffic analysis system will be very useful for network administrator to monitor packet flow and also to plan for the future. In this paper we focused on the flow analysis and flow control of packets generated by network topology. When we monitor a large volume of traffic data for detailed statistics, it is not easy to handle huge traffic with single server. As large dataset is required for matching computing and storage resources, Scalable Internet traffic measurement and analysis is difficult So we introduced Hadoop, an open-source computing platform of distributed file system, that has become a popular infrastructure for massive data analytics because it facilitates scalable hive data processing and storage services on a distributed computing system. The future work will show about the various problems causing the congestion in the network. It will also contain methodologies that must be implemented in order to avoid congestion in the network for the big data using Hadoop tool.

ACKNOWLEDGMENT

The completion of our project brings with it a sense of satisfaction, but it is never complete without them those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is the great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

REFERENCES

[1] Y. Lee and Y. Lee, "Toward scalable internet traffic measurement and analysis with Hadoop", ACM SIGCOMM Computer Communication Review, Volume 43, Number 1, January 2013.
[2] J. Liu, F. Liu, and N. Ansari, "Monitoring and Analysing Bi Traffic Data of a Large-Scale Cellular Network with Hadoop", IEEE Network July/August 2014.
[3] Hadoop, http://hadoop.apache.org .
[4] Chakchai Soin, "A Survey of Network Traffic Monitoring and Analysis Tools", soin@ieee.org,12 August 2014..
[5] http://hadoop.apache.org/docs/r1.0.4/hdfs_user_guide.html.
[6] Maheen Hasib and John A. Schormans, "Limitations of Passive & Active Measurement Methods in Packet Networks".
[7] Tcpdump, http://www.tcpdump.org .
[8] Wireshark, http://www.wireshark.org .
[9] L. Ibrahim, R. Hassan, K. Ahmad, A. Asat, "A study on improvement of Internet Traffic Measurement and Analysis Using Hadoop System" The 5th International Conference on Electrical Engineering and Informatics 2015 August 10-11, 2015, Bali, Indonesia.
[10] http://www.netresec.com/
[11] Dave Plonka Flow Scan, "A Network Traffic Flow Reporting and Visualization Tool", 12, May 2014.
[12] "Traffic Analysis with Netflow Whitepaper", 2005.
[13] Cisco Systems, "Simple Network Management Protocol", Internetworking Technologies Handbook, Chpt 56,1992-2006. http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/snmp.htm .
[14] Y. Lee, W. Kang, and Y. Lee, "A Hadoop-Based Packet Trace processing Tool", Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6613 LNCS, pp. 51–63, 2011.
[15] Y. Lee, W. Kang, and H. Son, "An Internet traffic analysis method with Map Reduce," 2010 IEEE/IFIP Netw. Open. Manage. Symp. Work., pp. 357–361, 2010.
[16] A W. Nagele, "Large-scale PCAP Data Analysis Using Apache Hadoop -RIPE Labs.", Available: https://labs.ripe.net/Members/wnagele/large-scale-pcap-data analy sis - using-apache-hadoop.
[17] Parekh B., Alka Patel," A Survey on Internet Traffic Measurement and Analysis", Computer Engineering and Intelligent Systems, Vol.6, No.4, 2015.
[18] Shankar Manikandan, Siddhartha Ravi," Big Data Analysis using Apache Hadoop", 978-1-4799-6541-0/14/$31.00 ©2014 IEEE.
[19] Alisha Cecil, " A Summary of Network Traffic Monitoring and Analysis Techniques".
[20] Yeonhee Lee and Youngseok Lee "Scalable NetFlow Analysis with Hadoop", Chungnam National University, Korea January 8, 2013 FloCon 2013.