Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.456 – 460

RESEARCH ARTICLE

Packet Flow Analysis and Congestion Control of Big Data by Hadoop

Dr. Mohammed Abdul Waheed[1], Mallappa Siragur[2], Lingaraj Veerappanavar[3]

[1]Department of Computer Science and Engineering, VTU RO Kalaburagi, India

[2]Department of Computer Science and Engineering, VTU RO Kalaburagi, India

[3]Department of Computer Science and Engineering, VTU RO Kalaburagi, India

dr.mawaheed@gmail.com [1] mallusiragur@gmail.com [2] lingarajdv15@gamil.com [3]

ABSTRACT-

Handling internet traffic in these is difficult, but the due to large scale datacenter network and explosive growth of internet traffic is hard to store and process the internet traffic on a single server. Hadoop has become a popular framework for massive data analytics. It facilitates scalable data processing and storage services on a distributed computing system consisting of commodity hardware. In this paper, I present a Hadoop based traffic analysis and control system, which accepts input from Wire shark (Log File), and output in form of summary which contains entire internet traffic details and I also implemented the congestion control algorithm to control the online network traffic in the internet

Keywords: Flow Clustering, Traffic Analysis, Software Defined network, Hadoop, OpenDaylight, Whireshark, Virtual Box.

I. INTRODUCTION

Analysis of big data is currently considered as an integral part of many computational and statisticaldepartments.As a result, novel approaches in data analysis are evolving on a daily basis. Thousands of transaction requests are handled and processed everyday by different websites associated with e-commerce, e-banking, e-shopping carts etc. The network traffic and weblog analysis comes to play a crucial role in such situations where Hadoop can be suggested as an efficient solution for processing the Network flow data collected from switches as well as website access-logs during fixed intervals.

The traffic data size is growing enormously but it isnecessary to extract the knowledge from this data by analyzing it. The networking devices and user devices are increasing rapidly with high performance which makes difficult for Internet Service Provider's (ISP's) to collect and analyze this traffic data. The ISP will need large infrastructure to store and analyze this data. But, again it leads to certain challenges such as fault-tolerant system, Performance, scalability, availability and many more which are faced by distributed system. Most of the time, ISP

will rely on Single High-Performance Server to analyze this traffic data. But, as the traffic data increases this method will become inefficient

Software defined network (SDN) is one of the method in computer networks which helps to govern the networks. Thisapproach helps in managing the networks by decoupling the system that compels outcome about where the traffic issent.For the traffic analysis of the Big data first thing we need to do is collect and measure the traffic data from varioussources. Big data is a tremendous collection of information which cannot be processed by traditional processingapplication. Big data refers to the volume variety and velocity of the data. It is not just dates, numbers, strings. It is alsoaudio, video, 3D data, unstructured text, social media and also log files. So it is a challenging task to measure andanalyze the Big data. Hadoop software library is a scaffold that uses simple programming techniques to process largedata sets. Yu [1] tells about software named Open Sketch that is designed for measuring traffic, this software splits thedata plane measurement from the control plane measurement. In this paper, system accepts input of large scale of trace file generated from traffic measurement tool like Wireshark, identifies flows running on the network from this trace file. The system further provides functions of detailed statistical analysis and characteristics mining. With characteristic information of individual flow, the controller can provide corresponding resources and service.

II. RELATED WORK

A lot of research is done to measure the performance of the internet traffic using Hadoop. ScsccJ. Shafer, S. Rixner, and Alan L [2]. Cox discuss about performance of distributed Hadoopfile system. Hadoop is most accepted framework for managing huge amount of data in distributed environment.

The core idea behinds clustering flows is clustering flows with their priori characteristics and statistical information. This method prevents suffering from influence of dynamic port and payload information. McGregor [7] proposes a methodology that can break packet header traces into clusters of traffic where each cluster has different traffic characteristics. Ermanevaluates two unsupervised clustering algorithms, namely K-Means and DBSCAN, and compares them to other previously used automatic classification algorithm. Hadoop makes use of user-level file system in distributed Internet traffic monitoring and analysis. Tcp-dump [11] is a powerful command-line packet analyzer with libpcap as a library for network traffic capture. Wireshark [10] is network protocol analyzer which is visually rich, powerful LAN analyzer that captures live traffic and stores it for offline analysis. Cisco IOS NetFlow Analyzer [10], GenieATM [8] provides facility for network traffic analysis. CoralReef [9] developed by CAIDA measures and analyzes passive Internet traffic data. Though all these are useful seems looking at first sight, most of these are run on single high performance server which is not capable of handling huge amount of traffic data captured at very high-speed links. Still there are multiple solutions but having some limitations. Hadoop built on MapReduce basics can be used to analyze web, text and log files. RIPE [4] does not consider the parallel-processing capability of reading the packet records from HDFS which leads to performance degradation. This study shows you the complete Internet traffic analysis system with Hadoop that can quickly process IP packets and NetFlow data.

III. TRAFFIC FLOW ANALYSIS AND CONTROL ARCHITECTURES

A. Overview

The Key components for the Flow Analysis using Hadoop consist of three main layers which include Data Exchange layer, Analysis layer and User Interface layer [6]. Fig 1 shows the key components required for Flow Analysis. The functions of the above 3 layers are described below:

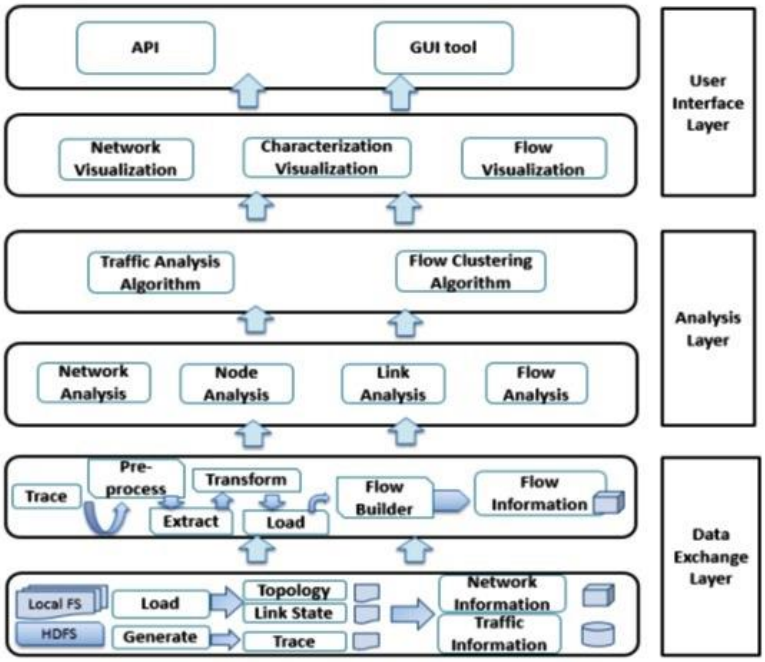

Fig. 1. Component Analysis

1***Data Exchange layer***: This layer implements HDFS (Hadoop Distributed File System) to store the information related to the Internet traffic. This layer is mainly concerned about the storage and it provides support to the other layers. In this layer preprocessing of the local file system is done. Here the network information and the traffic information are extracted from the packets which are got from the network.

2 ***Analysis layer:***This layer focuses of the internet traffic analysis and its management. In this layer multiple types of analysis are done. In this layer network analysis, node analysis, link analysis and flow analysis are done. Analysis layer also implements various algorithms needed for the flow analysis.

3***User Interface layer:***In this the user can interact with the system. The system will display graphical images to the user so the user can better understand the flow analysis. This layer implements few API and GUI tools for the better communication purpose.

B. System Architecture

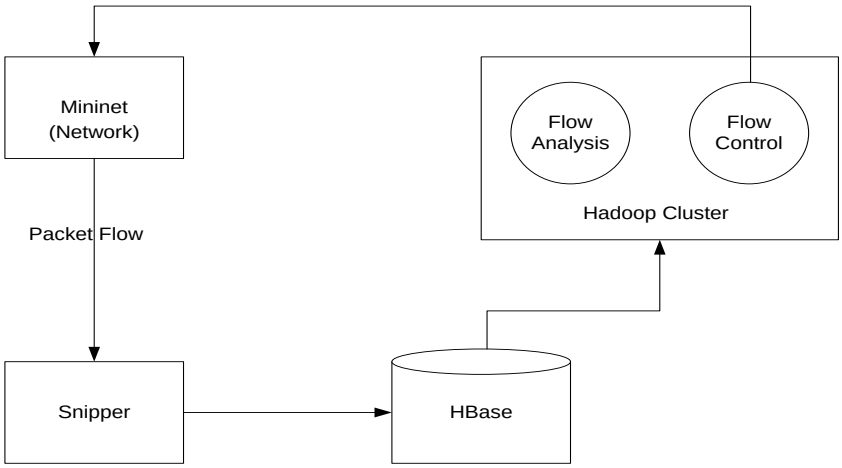

Fig. 2. Flow Analysis

Mininet: It is a network used to flow the packets. And flow control operation done accordingly with this module.

Snipper: It is a tool used to capture the snapshot on the dektop and forward it to the HBase

HBase: It is a database table used to store the snaps captured.

Hadoop Cluster: Clustering taking place in two processes Flow Analysis and Flow Control.

Flow Analysis: This class is used to analyze the traffic and get the flow information.

Flow Control: This class is used to do controlling of flow and get information of the same.

C. Flow Control Algorithm

As by looking the below the flows cannot be handled through single system if huge no of traffic is came then it cannot be handled so we planned to control the flows with in the network through this lot of congestion in the network packet can be avoided. In the below figure 3 show the entire network how the flows can be able control. I have implemented thecongestion control algorithm to control the flows from source host to destination hosts.If the packet is of byes exceeds above some range then the path form host -1 to host-2 can be changed orelse the packet bytes doesn‟t exceeds the range the old path form host-1 to host-2 can be used for packettransmission.Here it will check for bytes I wrote the algorithm that will handle control on only bytes.

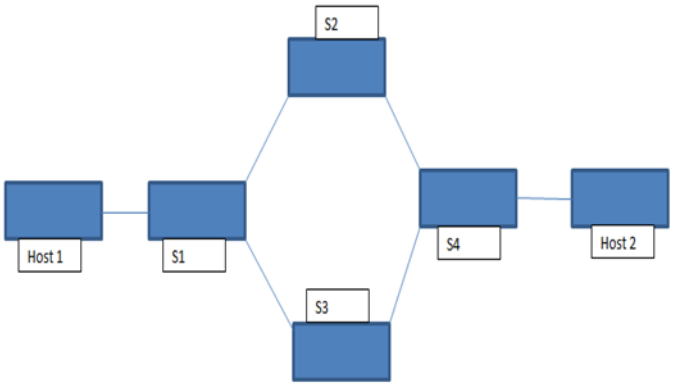

Fig. 3. Network Topology

As follows

1. Set up the network topology with two paths
2. Set one path with low bandwidth and other path with high bandwidth links
3. analyze flows between the hosts
4. Collect the number of packets exchanged between the host in terms of bytes
5. If no_of_bytes> threshold, switch to high bandwidth path by increasing the priority of the links.
6. If no_of_bytes< threshold, switch to low bandwidth path by decreasing the priority of the links.

IV. CONCLUSION AND FUTURE WORK

The traffic flow identification system will be very useful for network administrator to monitor faults and also to plan for the future. In this paper we focused on the flow analysis and flow control of packets of Trace files or log files generated by network topology and we gave a detailed analysis on how the packets are switched to low to high bandwidth vice versa when number of packets greater than threshold and lesser than threshold respectively.

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 bigdata using Hadoop tool

REFERENCES

[1] A. W. Moore and K. Papagiannaki, "Toward the accurate identificationof network applications," in Passive and Active Network Measurement. Springer, 2005, pp. 41–54

[2] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic classification on the fly," ACM SIGCOMM Computer Communication Review, vol. 36, no. 2, pp. 23–26, 2006.

[3] Scscc J. Shafer, S. Rixner, and Alan L. Cox, "The Hadoop Distribution Filesystem: Balancing Portability and Performance", in Proceedings of the 10th ACM SIGCOMM conference on Internet measurement ACM 2010.

[4] RIPE Hadoop PCAP, https://labs.ripe.net/Members/wnagele/large-scale-pcapdata-analysis-using-apache-hadoop, Nov. 2011

[5] M. Yu, L. Jose, and R. Miao, "Software defined traffic measurement with opensketch," in Proceedings 10th USENIX Symposium on Networked Systems Design and Implementation, NSDI, vol. 13, 2013.

[6] Apache Hadoop Website, http://hadoop.apache.org/

[7] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," in Passive

[8] GenieATM, http://www.genienrm.com

[9] CAIDA CoralReef Software Suite,

[10] perfSONAR. http://www.perfsonar.net.

[11] Wireshark, http://www.wireshark.org

[12] Tcpdump, http://www.tcpdump.org.

