A Survey on Internet Traffic Measurement and Analysis

Parekh Nilaykumar B.
Department of CS&E, Governmernt Engineering Collage,Modasa, Aravalli,Gujarat,India
Tel: +91 9913885777 E-mailnilaybparekh@yahoo.in

Prof. Alka J. Patel
Department of CS&E,Governmernt Engineering Collage,Modasa, Aravalli,Gujarat,India
Tel: +91 9427402494 E-mail: alkapatel227@gmail.com

Abstract:

As the number of Internet users increasing rapidly in this world, Internet traffic is also increased. In computer network traffic measurement is the process of measuring the amount and type of traffic on a particular network. Internet traffic measurement and analysis are mostly used to characterize and analysis of network usage and user behaviour, but faces the problem of scalability under the explosive growth of Internet traffic and high speed access. It is not easy to handle Tera and Pera-byte traffic data with single server. Scalable Internet traffic measurement and analysis is difficult because a large dataset requires matching commutating and storage resources. To analyse this traffic multiple tools are available. But they do not perform well when the traffic data size increase. As data grows it is necessary to increase the necessary infrastructure to process it. The distributed File System can be used for this purpose, but it has certain limitation such as scalability, availability and fault tolerant. Hadoop is popular parallel processing framework that is widely used for working with large datasets and it is an open source distributed computing platform having MapReduce for distributed processing and HDFS to store huge amount of data. In future work we will present a Hadoop-based traffic monitoring system that perform a multiple types of analysis on large amount of internet traffic in a scalable manner

Keywords- Traffic monitoring, Hadoop, MapReduce, HDFS, NetFlow.

1. Introduction

Network monitoring and measurement have become more and more important in a modern complicated network. In the past, administrators might only monitor a few network devices or less than a hundred computers. The network bandwidth may be just 10 or 100 Mbps (Megabit per second) ; However, now administrators have to deal with not only higher speed wired network (more than 10 Gbps (Gigabit per sec) and ATM (Asynchronous Transfer Mode) network) but also wireless networks. They need more sophisticated network traffic monitoring and analysis tools in order to maintain the network system stability and availability such as to fix network problems on time or to avoid network failure, to ensure the network security strength, and to make good decisions for network planning [1]. When a network failure occurs, monitoring agents have to detect, isolate, and correct malfunctions in the network and possibly recover the failure. Commonly, the agents should warn the administrators to fix the problems within a minute. With the stable network, the administrators' jobs remain to monitor constantly if there is a threat from either inside or outside network. Moreover, they have to regularly check the network performance if the network devices are overloaded. Before a failure due to the overload, information about network usage can be used to make a network plan for short-term and long-term future improvement.

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 [9]. 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 he assumed that he “owns” all networks [2].

In the network traffic measurement there are mainly two challenges like 1) Flow statistics computation time and 2) Single node failure. To address these challenge, I want to implement the internet traffic measurement and analysis using MapReduce programming model of Hadoop framework. Apache Hadoop is an open source software frame work for storage and large scale processing of Netflow datasets. MapReduce is a programming model and an associated implementation for processing and generating large datasets that is amenable to a broad variety of real-world tasks.

2. Literature Survey

2.1 Applications

This proposed will provide a new approach to measuring and analysis Internet traffic which is based on MapReduce Paradigm. This new approach will try to improve the computational time, more fault tolerance of system and will handle or deal with Bigdata during internet traffic measurement and analysis.

2.2 Related Work

This chapter will give the idea of how different network traffic tools is working on their environment. After that we will discuss and classify existing approaches for Internet traffic measurement and analysis with their advantages and limitation. This chapter is written according to literature survey done by us for preventing disclosure of sensitive association rules from unauthorized access.

Internet traffic measurement and analysis has long been used to characterize network usage and user behaviours. 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.

Tool Name	Tool Reviews
Cisco NetFlow	Cisco NetFlow easily monitors flows passing through routers and switches without observing each packet. Unfortunately its format is not open and it has been designed only for IPv4 network monitoring [1].
Wireshark	Wireshark isn't an intrusion detection system. It will not warn you when someone does strange things on your network that he/she isn't allowed to do and Wireshark will not manipulate things on the network [7].
Tcpdump	Tcpdump tool has some limitation like- this tool can only focus on what it find, mean Tcpdump can report on only what it finds in the packets. If IP address is forged in the packets, tcpdump has no ability to report anything else [7].
Pandora FMS	If you want to keep on your eyes on services, applications and communication, this tool is best for this purpose but Pandora FMS get struggled when we want to manage large network environment [15].
Angry IP Scanner	Angry IP scanner scans IP address in its port and finding live hosts and providing you information about them. There are no obvious drawbacks to mention but Sometimes Angry IP Scanner cannot detect open ports and will consider them as filtered [16].
Network Miner	The main purpose of Network Miner is data collection for future analysis (forensic evidence analysis) rather than collecting data regarding the traffic on the network. Information are grouped by host rather than by packets or frames [17].

3. Background Study

3.1 Big Data

Big data is a term for massive data sets, a large amount of data available in complex structured or no structure form. These vast amounts of data are generated by social media and networks, scientific instruments, mobile devices, sensor technology an networks. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics.If the data which is beyond to the storage capacity and which is beyond to the processing power that data is calling BigData [11].

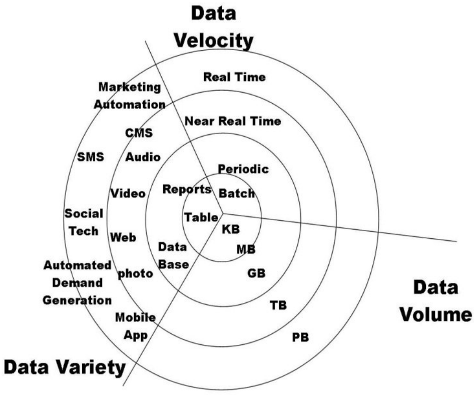

Figure 3.1: The three V's of Big data [19]

3.2 Apache Hadoop Framework

Apache Hadoop [23] is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. The Apache™ Hadoop® [24] project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop framework includes following modules:

- Hadoop Common: It having utilities that support the other hadoop module.
- Hadoop distributed File System (HDFS): a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
- Hadoop YARN: A framework for job scheduling and cluster resource management.
- Hadoop MapReduce: a programming model for large scale data processing.

3.2.1 Hadoop Distributed File System

HDFS stands for "Hadoop Distributed File System" and HDFS is highly scalable file system. HDFS supports parallel reading and processing data. HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

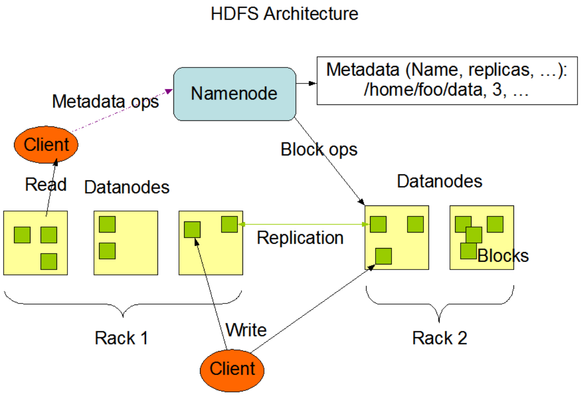

Figure 3.2 HDFS Architecture [12]

3.2.2 MapReduce

Map/reduce is a special form of such a DAG (Direct acyclic Graph) which is applicable in a wide range of use cases and it's a programming paradigm for processing large datasets in distributed environment.It is organized as a "map" function which transform a piece of data into some number of key/value pairs. Each of these elements will then be sorted by their key and reach to the same node, where a "reduce" function is use to merge the values (of the same key) into a single result.

```
map (input_record)
{
        ……
        emit (k1, v1)
        ……
        emit (k2, v2)
        ……
}
```

```
reduce (key, values)
{
     aggregate = initialize()
     while (values.has_next)
      {
         aggregate = merge(values.next
       } collect(key, aggregate)
 }
```

3.3 Why Hadoop??

Hadoop is an open source framework given by apache software foundation for storing huge dataset and for processing huge dataset with the cluster of commodity hardware. So basically Hadoop is for storing and processing bigdata.

4. Design of Experiments

For the experiments we used, Wireshark version 1.12.4 ,Operating System: 64-bit Ubuntu 14.04 LTS with Intel Core i5 CPU @ 2.30 GHz × 4 and 3 GB of RAM.

Also we were took the dataset in .pcap file format from the Government Engineering College, Modasa Gujarat India.

Dataset type: .pcap file

Number of packets: 305916, 22189601

4.1 Experiment and Result

Experiment: Protocol Summarisation analysis for pcap file in Wireshark tool.

In this experiment we first took pcap file as input, which contains 305916 packets. We feed that packets in one of the famous network analysis tool "Wireshark" and this tool took 5.727 seconds to perform protocol summarisation analysis. Then we took another pcap file, which contain 22189601 packets in "wireshark" tool and this tool took 1.51 minutes means 111 seconds for Protocol
Summarisation Analysis.

No of Packets & Size of file	Protocol Summarisation analysis Time in Wireshark (sec)
305916 (250 MB)	5.727
22189601 (1 GB)	111

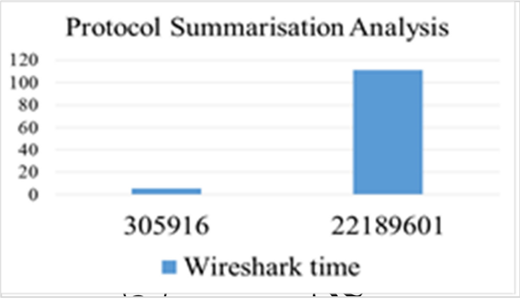

Figure: 4.1 Results of small experiments

4.2 Problem Gap

From the protocol summarisation experiment, we can easily understand that when data is small in size, wireshark tools result is good , but as we can see that when data goes big, the analysis result is time consuming, so we have come up with another approach for same problem, that is Hadoop Technology.

Hadoop is an open source framework given by apache software foundation for storing huge dataset and for processing huge dataset with the cluster of commodity hardware. So basically Hadoop is for storing and processing bigdata.

5. Conclusion

Internet traffic measurement and analysis have been usually performed on a high performance server that collects and examine packets. When we monitor a large volume of traffic data for detailed statistics, it is not easy to handle Tera or Pera-bytes traffic with single server. Scalable Internet traffic measurement and analysis is difficult because a large data set require matching computing and storage resource. So we introduced Hadoop, an open-source computing platform of MapReduce 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.

Reference

[1] Chakchai SoIn, soin@ieee.org: *A Survey of Network Traffic Monitoring and Analysis Tools*, 12 *August 2014*
[2] Maheen Hasib and John A. Schormans: *Limitations of Passive & Active Measurement Methods in Packet Networks*

[3] QIAO Yuan-yuan1, LEI Zhen-ming, YUAN Lun2, GUO Min-jie: Offline traffic analysis system based on Hadoop 2003

[4] Ryoichi Kawahara, Keisuke Ishibashi, Toshiyuki Hirano, Hiroshi Satio, Hisaki Ohara, Daisuke Satoh, Shochiro Asano , Jun Matsukata:*Traffic Measurement and analysis in an ATM-based internet backbone* 11 January 2001

[5] Xiaofeng Zhou, Milenko Petrovic, Tom Eskridge, Marco Carvalho, Xi Tao: *Exploring Netow Data using Hadoop,* 2014 Ase Bigdata/Socialcom/Cybersecurity Conference, Stanford University, May 27-31, 2014

[6] Network Traffic Analysis Using Cisco NetFlow: *White paper, ©2007 SolarWinds, Inc. www.solarwinds.com, 866.530.8100*

[7] Pallavi Asrodia, Hemlata Patel: *Analysis of Various Packet Sniffing Tools for Network Monitoring and Analysis*, International Journal of Electrical, Electronics And Computer Engineering 1(1): 55-58(2012)

[8] Havar Aambo Fosstveit, Erlend Hamberg, Espen Auran Rathe, Per Oyvind Stadheim*: Making the Wireshark Network Analyzer Eat Less Memory*, 18 November 2009

[9] Network Traffic Measurement: http://en.wikipedia.org/wiki/Network_traffic_measurement

[10] MapReduce framework: http://d2i.indiana.edu/hmr

[11] BigData Wikipedia: http://en.wikipedia.org/wiki/Big_data

[12] Hadoop: http://Hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[13] MapReduce: http://architects.dzone.com/articles/how-Hadoop-mapreduce-works

[14] NetFlow Architecture: http://en.wikipedia.org/wiki/NetFlow

[15] Pandora FMS: http://en.wikipedia.org/wiki/Pandora_FMS

[16] Angry IP Scanner: http://angryip.org

[17] Microsoft Network Monitor: http://en.wikipedia.org/wiki/Microsoft_Network_Monitor

[18] Romain Fontugne,Johan Mazel, Kensuke FukudaHashdoop: A MapReduce Framework for NetworkAnomaly DetectionCERN – European organization for nuclear Research 2014 IEEE INFOCOM Workshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data,2014

[19] BigData : http://velvetchainsaw.com/2012/07/20/three-vs-of-big-data-as-applied-conferences

[20] Dave Plonka FlowScan: A Network Traffic Flow Reporting and Visualization Tool 12, May 2014

[21] http://www.caida.org/data/passive/passive_2010_dataset.xml

[22] Network Analysis Using Wireshark Cookbook, by Yoram Orzach , Packet Publication, BIRMINGHAM – MUMBAI, Copyright © 2013 Packt Publishing

[23]Apache Hadoop, December 2014: http://en.wikipedia.org/wiki/Apache_Hadoop

[24] Welcome to Apache Hadoop, December 2014: http://hadoop.apache.org/