Network Flow Analysis using Big Data

Yogesh Dorbala, Kishore Rajendra

September 2, 2014

Abstract

In our project, we plan to use Hadoop to analyze very large data-sets of Network- Flow data for different process.There are many new languages and databases to support Big data framework Hadoop. We would exlpore to find the best suitable tools and languages for our project. We will use different methods described in other papers to accurately analyze the network flow. Using Big data analysis, we have the power and capability to analyze very large sets of data in very less time.

1 Introduction

The rate of data creation has increased so much that 90 of the data in the world today has been created in the last two years alone. This acceleration in the production of information has created a need for new technologies to analyze massive data sets. The term Big Data refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies.

Big Data Analytics is a paradigm where large data sets are split into small ones and given to worker nodes to compute and all the results are appended to give the result.This has a huge advantage as data with variety, velocity and volume can be managed easily. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing.

Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics. Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion.

Organization The rest of the paper is organized as follows: Hadoop description is given in section II. In the section III, we provide literature survey on the usage of MapReduce with Hadoop. In section IV, we describe the exisiting security measures against DDoS.Finally, Conclusion and future work are included in section VI.

2 Hadoop

The Apache Hadoop is a framework that allows for the storing large data sets which are distributed across clusters of computers using simple programming models and written in Java to run on a single computer to large clusters of commodity hardware computers.

It is derived from papers published by Google and incorporated the features of the Google File System (GFS) and MapReduce paradigm which are named as Hadoop Distributed File System and Hadoop MapReduce.

2.1 Key Characteristics

1. **Scalable**– New nodes can be added as needed, and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.
2. **Economical**– Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.
3. **Flexible**– Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.
4. **Reliable**– When you lose a node, the system redirects work to another location of the data and continues processing without missing a beat

2.2 Motivation behind using Hadoop

With development of new technology and devices, rate of collection of data is increasing very rapidly. The volume of business data worldwide, across all companies, doubles every 1.2 years, according to estimates.
Handling such large databases, datasets and queries is impossible using traditional technologies. This is where Big data is way ahead of other technologies. It leverages the parallel computing power and eliminates the need for high-end devices by using commodity machines, reducing the cost of computation.

According to Cisco estimates,

1. Global IP traffic willreach 1.1 zettabytes per year or 91.3 exabytes (one billion gigabytes) per month in 2016. By 2018, global IPtrafficwill reach 1.6 zettabytes per year, or 131.6 exabytes per month.
2. Global Internet traffic in 2018 will be equivalent to 64 times the volume of the entire global Internet in 2005
3. Globally, mobile data traffic will increase 11-fold between 2013 and 2018.
4. Global mobile data traffic grew 81 percent in 2013. Global mobile data traffic reached 1.5 exabytes per month atthe end of 2013, up from 820 petabytes per month at the end of 2012.

A few Examples involving huge data.

1. Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second.The data flow would exceed nearly 500 exabytes per day. This is equivalent to 500 quintillion (5*1020) bytes per day, almost 200 times higher than all the other sources combined in the world.

2. Decoding the human genome originally took 10 years to process, now it can be achieved in less than a day : the DNA sequencers have divided the sequencing cost by 10,000 in the last ten years, which is 100 times cheaper than the reduction in cost predicted by Moore's Law.

3. eBay.com uses two data warehouses at 7.5 petabytes and 40PB as well as a 40PB Hadoop cluster for search, consumer recommendations, and merchandising.

4. Amazon.com handles millions of back-end operations every day, as well as queries from more than half a million third-party sellers. The core technology that keeps Amazon running is Linux-based and as of 2005 they had the world's three largest Linux databases, with capacities of 7.8 TB, 18.5 TB, and 24.7 TB

3 Literature Survey

Big Data analytics can be leveraged to improve information security and situational awareness. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

The report explores how to use Big data Analytics to solve current problems with having to analyze Terabytes of data through the help of research paper publications.

3.1 Detecting Botnets

Botnets are a major threat of the current Internet. Understanding the novel generation of botnets relying on peer to-peer networks is crucial for mitigating this threat. Nowadays, botnet traffic is mixed with a huge volume of benign traffic due to almost ubiquitous high speed networks. Such networks can be monitored using IP flow records but their forensic analysis form the major computational bottleneck.

The paper *"BotCloud: Detecting Botnets Using MapReduce"* describes a scalable method for detecting P2P botnets regarding the relationships between hosts. The evaluation shows a good detection accuracy and a good efficiency based on a Hadoop cluster.

3.2 Detecting DDoS Attacks

Recent distributed denial-of-service (DDoS) attacks have demonstrated horrible destructive power by paralyzing web servers within short time. As the volume of Internet traffic rapidly grows up, the current DDoS detection technologies have met a new challenge that should efficiently deal with a huge amount of traffic within the affordable response time. This paper devised a DDoS anomaly detection method on Hadoop that implements a MapReduce-based detection algorithm against the HTTP GET flooding attack.

In the paper *"Detecting DDoS Attacks with Hadoop"*, the focus was on a scalability issue of the anomaly detection. The paper also introduced a Hadoop-based DDoS detection scheme to detect multiple attacks from a huge volume of traffic. This method leverages Hadoop to solve the scalability issue by parallel data processing.

3.3 Brute-force Attack Detection

The paper *"Flow-based Brute-force Attack Detection"* presents several methods for the detection of brute-force attacks based on the analysis of network flows and discusses their strengths and shortcomings. It also demonstrates the fragility of some methods by introducing detection evasion techniques.

This chapter gave an overview about current research in the field of flow-based attack and anomaly detection. It also summarized state of the art concepts for attack detection, especially brute-force ones, and concluded the chapter by discussing evasion strategies as well as the limitations inherent to the process of detecting attacks in network ows.

4 Distributed Denial of Service

DDoS is a type of DoS attack where multiple compromised systems which are usually infected with a Trojan are used to target a single system causing a Denial of Service (DoS) attack. Victims of a DDoS attack consist of both the end targeted system and all systems maliciously used and controlled by the hacker in the distributed attack.

Defence Measures against DDoS

1. System security mechanisms increase the overall security of the system, guarding against illegitimate accesses to the machine, removing application bugs and updating protocol installations to prevent intrusions and misuse of the system.

2. Protocol security mechanisms address the problem of bad protocol design. Many protocols contain operations that are cheap for the client but expensive for the server. Such protocols can be misused to exhaust the resources of a server by initiating large numbers of simultaneous transactions.

3. Ingress Filtering, proposed by Ferguson and Senie, is a restrictive mechanism to drop traffic with IP addresses that do not match a domain prefix connected to the ingress router. Egress filtering is an outbound filter, which ensures that only assigned or allocated IP address space leaves the network.

4. Misuse detection identifies well-defined patterns of known exploits and then looks out for occurrences of such patterns. Several popular network monitors perform signature-based detection, such as CISCO'S NetRanger, NID, Realsecure, Snort.

5 Approach

5.1 Statistical Analysis

Using Hadoop's MapReduce paradigm,the map function filters non-HTTP GET packets and generates key values of server IP address, masked timestamp, and client IP address. The reduce function summarizes the number of URL requests, page requests, and server response between a client and a server. Finally, the algorithm aggregates values per server. When total requests for a specific server exceeds the threshold, the records are marked as attackers.

A Simple example:
We have a .pcap file generated from Wireshark.We will do offline-analysis in this example. After running our program for Packet Count, we get the following output sorted in ascending order or IP Address

74.125.236.36 - 40
134.170.188.221 - 58
173.252.110.27 - 182

Here first column is Source IP Address and second column is the number of packets from that IP Address. Similarly we will collect other data like src and destination port number, total hops,payload

length, flags, timestamp and other information. At the end we can calculate how many connections were made with the host and gather data like average payload length for that session. This is particularly helpful in detecting attacks as those sessions have slightly different properties which deplete the resources of the host.

5.2 Hop-count Method

Each packet has a TTL (Time to live) value.Whenever it passes through a router, TTL value is reduced by 1. THis is done so that packets whose destination is not reachable do not overload the network. When a packet is created, it has some TTL value depending on the Operating System.

OS Version	"safe"	tcp_ttl	udp_ttl
AIX	n	60	30
DEC Pathworks V5	n	30	30
FreeBSD 2.1R	y	64	64
HP/UX 9.0x	n	30	30
HP/UX 10.01	y	64	64
Irix 5.3	y	60	60
Irix 6.x	y	60	60
Linux	y	64	64
MacOS/MacTCP 2.0.x	y	60	60
OS/2 TCP/IP 3.0	y	64	64
OSF/1 V3.2A	n	60	30
Solaris 2.x	y	255	255
SunOS 4.1.3/4.1.4	y	60	60
Ultrix V4.1/V4.2A	n	60	30
VMS/Multinet	y	64	64
VMS/TCPware	y	60	64
VMS/Wollongong 1.1.1.1	n	128	30
VMS/UCX (latest rel.)	y	128	128
MS WfW	n	32	32
MS Windows 95	n	32	32
MS Windows NT 3.51	n	32	32
MS Windows NT 4.0	y	128	128

In this method, we store a table with IP-hopcount entries. Source-IP Address and number of hops for that IP. We ping recently contacted IP Addresses regulary and record how many hops packets take.When a packet enters our network, we check its hopcount entry. But one problem with this approach is, packets may not take the same route always as some devices or networks might go down, due to which route will change.

We propose to check for first 1 or 2 IP Addresses from source. Because the source's ISP remains same for some period,lets say it is IP1. we can conclude if a packet is valid by checking the first hop-IP. If it is same as IP1,

Example:
Lets say usual path from src to host is:

src - IP1 - IP2 - IP3 —- IP10 - host (no. of hops == 12)

Due to some problems, router with IP2 is down, the new path is:
src - IP1 - IPnew1 - IPnew2 - IP3 —- IP10 - host (no. of hops != 13)

But it is valid packet because IP2 is down, and thats why TTL value will be different now.

6 Conclusion

Report on the Network flow analysis using BigData taking into consideration all the factors governing the flow. After collecting real-data for a network flow, we will analyse and detect any anomolies or attacks. If required, we may have to generate attacks by ourselves/ to verify our system. The BTP end deliverables are methods to prevent or detect attacks using Big-Data framework Hadoop.

7 References

"BotCloud: Detecting Botnets Using MapReduce", by Jerome Francois, Shaonan Wang, Walter Bronzi, Radu State, Thomas Engel. *"Detecting DDoS Attacks with Hadoop"*, by Yeonhee Lee, Youngseok Lee. *"Flow-based Brute-force Attack Detection"*, by Martin Drasar and Jan Vykopal.

"DDoS Attacks and Defence Mechanisms: A Classification" Christos Douligeris and Aikaterini Mitrokotsa

"A Taxonomy of DDoS Attacks and DDoS Defense Mechanisms" Jelena Mirkovic, Janice Martin and Peter Reiher

"Denial of Service Attacks" by Qijun Gu,Peng Liu. *"Network Security and DoS Attacks"* by Sílvia Farraposo,Laurent Gallon,Philippe Owezarski.

"Implementing Pushback: Router-Based Defense Against DDoS Attacks" by John Ioannidis,Steven M. Bellovin.