DETECTION OF FLOODING BASED DOS ATTACK ON HADOOP DATANODE

Silky kalra, Anil Lamba

Abstract—Denial of service (DOS) attacks present an internet-wide threat. SYN flooding is one of the DOS attacks that degrade the performance of the system. It is the most powerful attack used by hackers to harm the organization. It causes service outages and loss of millions, depending on the duration of attack. SYN flood DOS attack involves sending too many SYN packets to the destination. The attack use up the resources and memory of the server. This can lead to the hang of the server machine. Dos attacks are a persistent problem for several reasons. First, they are one of the earliest attacks to perform, and they attain quick results. They are the most common attacks that an administrator with a system that's always connected to the internet can expect. Well it is an attempt to make a system or server unavailable for legitimate users and, finally, degrades the service. This is accomplished by flooding the server's request queue with fake requests. After this, server will not be able to interact with the requests of legitimate users.

This paper focused on dos attack and how to overcome it using “HADOOP”. For this here an open source tools and softwares are used. We proposed a technique to detect flooding based attack at hadoop datanode and analysed the working of hadoop distributed file system. We have also shown hadoop’s effectiveness in attack scenario, discussed various motivation for deployment

Index Terms—Denial-of-Service(DOS), Datanode, Hadoop, hadoop distributed file system, Mapreduce,

I. INTRODUCTION

SYN flooding is one of the typical DoS attacks that exploit normal TCP connections between a client and a target web server by sending too many SYN packets to the destination server. This can lead to a crash or hang of server machine. As the volume of Internet traffic increases explosively year after year, the Intrusion Detection Systems (IDSes) have faced the issue on how to assure both scalability and accuracy of analyzing the DoS attack from these huge volume of data. In recent years, several approaches have been proposed to solve this issue. Dimensionality reduction methods such as Principal Component Analysis (PCA) enables to classify large volume of traffic by separating the normal behavior from anomalies [8].

However, these schemes usually require too excessive computing cycles to apply to actual systems. Sketch-based studies focus on memory efficiency by utilizing hash tables. Though Liu et al. [9] proposed a two-level sketch approach to reduce memory consumption and searching complexity while boosting accuracy, their technique still needs sufficient memory space and complex computation. Hadoop is an open-source distributed cluster plat-form that includes a distributed system, HDFS and the programming model, MapReduce.

The Iranian Cyber Army: On December 17, 2009, attackers replaced the front page of Twitter.com with an image of the Iranian flag along with text including: “This site has been hacked by the Iranian Cyber Army.” The attackers did not actually gain access to Twitter's servers, but instead changed the twitter.com domain name to point to a different IP address (the IP address of the machine hosting the "hacked by...") page. Twitter took down its home page entirely within minutes and twitter.com remained down for a couple of hours.[8]

The attacks on the major Web sites began in early February 2000, with the first major attack being on Yahoo! On February 7 [10]. The surprise attack took the Yahoo! Site down for more than three hours. It was based on the Smurf attack, and most likely, the Tribe Flood Network technique. At the peak of the attack, Yahoo! was receiving more than one gigabit per second of data requests.

Anonymous and “Operation Titstorm” - In February, 2010, a group of people loosely connected through Internet forums calling itself “Anonymous” executed a DDoS attack against the Australian Parliament’s website. The attack took down the site for two days. On the same day that Anonymous attacked the parliament’s website, the group also defaced the Prime Minister’s website, briefly replacing the front page with pornographic images. The attack was termed “Operation Titstorm” by its organizers, referring to a mandatory Internet filtering policy proposed by Australia’s ruling party designed in part to counter pornography.

Hadoop has created a lot of interest in large-scale analytics (the MapReduce part of Hadoop). This kind of “divide and conquer” algorithm methodology has been used for numerical analysis for many years as a way of dealing with problems that were known to be bigger than the biggest machine available. MapReduce and its open-source implementation Hadoop were originally optimized for large batch jobs such as web index construction. However, another use case has recently emerged sharing a MapReduce cluster between multiple users, which run a mix of long batch jobs and short interactive queries over a common data set.
As shown in fig1 MapReduce is an elegant way of structuring this kind of algorithm that isolates the analyst/programmer from the specific details of managing the pieces of work that get distributed to the available machines, as well as an application architecture that doesn’t depend on any specific structuring of the data.

II. RELATED STUDY

In [1], Prashant Chauhan, Abdul Jhummarwala, Manoj Pandya in December, 2012 provided an overview of Hadoop. This type of computing can have a homogeneous or a heterogeneous platform and hardware. The concept of cloud computing and virtualization has derived much momentum and has turned a more popular phrase in information technology. Many organizations have started implementing these new technologies to further cut down costs through improved machine utilization, reduced administration time and infrastructure costs. Cloud computing also confronts challenges. One of such problem is DDoS attack so in this paper author will focus on DDoS attack and how to overcome from it using honeypot. For this here open source tools and software are used. Typical DDoS solution mechanism is a single host oriented solution that meets scalability.

In [2], Jin-Hyun Yoon, Ho-Seok Kang and Sung-Ryul Kim, in 2012, proposed a technique called "triangle expectation" is used, which works to find the sources of the attack so that they can be identified and blocked. To analyze a large amount of collecting network connection data, a sampling technique has been used and the proposed technique is verified by experiments.

In [3], B. B. Gupta, R. C. Josha, Manoj Misra, in 2009, the main aim of this paper is First to demonstrate a comprehensive study of a broad range of DDoS attacks and defense methods proposed to fight with them. This provides a better understanding of the problem, current solution space, and future research scope to fight down against DDoS attacks. Second is to offer an integrated solution for entirely defending against flooding DDoS attacks at the Internet Service Provider (ISP) level.

In [4], Yeonhee Lee, Youngseok Lee, in 2011 proposed a novel DDoS detection method based on Hadoop that implements an HTTP GET flooding detection algorithm in MapReduce on the distributed computing platform.

In [5], Matei Zaharia, Dhruba Borthakur, Joydeep Sen Sarma, Khaled Elmeleegy, Scott Shenker, Ion Stoica, in April 2009, provided an overview of Sharing a MapReduce cluster between users. It is attractive because it enables statistical multiplexing (lowering costs) and allows users to share a common large data set. They evolved two simple techniques, delay scheduling and copy-compute splitting, which improve throughput and response times by factors of 2 to 10. Although we concentrate on multi-user workloads, our techniques can also increase throughput in a single-user, FIFO workload by a factor of 2.

In [6], Radheshyam Nanduri, Nitesh Maheshwari, Reddy Raja, Vasudeva Varma, in 2011, proposed an approach which attempts to hold harmony among the jobs running on the cluster, and in turn minify their runtime. In their model, the scheduler is made reminful of different types of jobs running on the cluster. The scheduler tries to assign a task on a node if the incoming task does not affect the tasks already running on that node. From the list of addressable pending tasks, our algorithm pick out the one that is most compatible with the tasks already running on that node. They bring up heuristic and machine learning based solutions to their approach and attempt to maintain a resource balance on the cluster by not overloading any of the nodes, thereby cutting down the overall runtime of the jobs. The results exhibit a saving of runtime of around 21% in the case of heuristic based approach and approximately 27% in the case of machine learning based approach when compared to Yahoo’s Capacity scheduler.

In [7], Dongjin Yoo, Kwang Mong Sim, in 2011, compare contrasting scheduling methods, evaluating their features, strengths and weaknesses. For settlement of synchronization overhead, two categories of studies; asynchronous processing and speculative execution are addressed. For delay scheduling in Hadoop, Quincy scheduler in Dryad and fairness constraints with locality improvement are addressed. In [8] Dileep Kumar Gupta, Abhishek Mishra et al, discussed the availability problem in the existing framework for e-Governance and also provide a better solution to solve availability problem in future framework for e–governance in cloud computing. They have proposed here a new modified Model by adding one filtering module in the existing algorithm. Basically, DoS attacks are used for two purposes. First is to consume the resources and second is to consume the bandwidth of network. In both cases, either resources or bandwidth of network are scarce. The most difficult part to defend against DoS attack is that, how to differentiate between normal traffic and malicious traffic? DoS attack has two solutions. It blocks the packets either from the port numbers or by the IP addresses. When blocking of packets is done by port number then it will
block all the packets coming from the particular port. For example, if we allow TCP packets to come into the network so that all UDP packets will drop and we cannot confirm that all TCP packets are coming from authenticated user so that we have used IP filtering mechanism to protect DoS attack.

III. PROPOSED FRAMEWORK

We have observed the effectiveness of hadoop in different attack scenario. Hadoop consist of two core components: the job management framework that handles the map and reduce tasks and hadoop distributed file system. We introduce the syn flooding attack with the help of code attached to hadoop and then captured it with wireshark. Datanode of HDFS receives the blocks of data and deletes the flooded blocks and a fair scheduler for better job management in which small adhoc query jobs can be executed with periodic jobs (for monitoring) in parallel that prevents the degradation in performance of distributed file system.

IV. METHODOLOGY

Step1: Flooding on hadoop datanode
Step2: Capturing the live traffic
Step3: copying that file to hadoop user
Step4: job assignment
Step5: map and reduce task
Step6: delete the flooded blocks of data
Step7: Collecting results

V. IMPLEMENTATION

We added a code for flooding at local host in c

```c
int main (void)
{
    int s = socket (PF_INET, SOCK_DGRAM, IPPROTO_TCP);
    char datagram[696], source_ip[32];
    struct iphdr *iph = (struct iphdr *) datagram;
    struct tcp hdr = struct tcp hdr (* datagram + sizeof (struct iphdr));
    struct sockaddr_in sin;
    struct pseud_headers ph;
    strcpy(source_ip, "192.168.1.2");
    sin.sin_family = AF_INET;
    sin.sin_port = htons(80);
    sin.sin_addr.s_addr = htonl ("127.0.0.1");
```

And then captured the traffic using wireshark. And then pass it to hadoop datanode.

VI. RESULTS

Datanode of hadoop distributed file system receives the blocks of data and deletes the flooded blocks.

---

Figure 2: Shows the capturing of traffic

Figure 3: Logs of datanode

Here job details are shown:
- Job_id is assigned to each job
- It maps and reduce tasks
VII. CONCLUSION

Detection of DoS attack is focused in this paper by Hadoop based DoS detection model. Flooding based DoS attack involves large number of packets sent to the hadoop within a short span of time. We have justified it with wireshark. We use a distributed detection system to efficiently detect these attacks at an early stage. MapReduce technique of Hadoop is used for distributing the analysis task to idle workers in the Hadoop cluster and gets that job done efficiently and accurately.

APPENDIX

A. List Of Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDFS</td>
<td>Hadoop distributed file system</td>
</tr>
<tr>
<td>DOS</td>
<td>Denial of service</td>
</tr>
</tbody>
</table>

B. TOOLS USED FOR IMPLEMENTATION

Hadoop
Wireshark

ACKNOWLEDGMENT

While preparing this article I got unlimited support and spirit from my family and friends. I am highly grateful to the Principal, Haryana Engineering College, Jagadhri for providing me the opportunity to carry out my research. I express my deep sense of gratitude to my Research Supervisor Er. Amit lamba, for his expert guidance, stimulating discussions as well as continued impetus throughout the period of this work. I would like to thank wholeheartedly Er. Anil Lamba, Chairperson, Department of CSE, HEC Jagadhri for encouraging solicited advice and all kind of help extended during the period of this work. I also express my gratitude to Er. Suveg Moudgil for his valuable and precious help during this research. Finally, I am thankful to all those people who are directly or indirectly related with this work. Above all I pay my regards to the Almighty for always having his hand on me.

REFERENCES

[2] Jin-Hyun Yoon, Ho-Seok Kang and Sung-Ryul Kim, Division of Internet and Media, Konkuk University, Seoul, Republic of Korea h2jhyyoon@gmail.com, hsriver@gmail.com, kimsr@konkuk.ac.kr , pp. 200-203
[3] B. B gupta. , Joshi, R. C. and Misra, Manoj(2009), 'Defending against Distributed Denial of Service

Yeonghee Lee, Changnam National University, Daejeon, 305-764, Republic of Korea, yhlee06@cmu.ac.kr


Jose Nazario, “Politically Motivated Denial of Service Attacks.”


Geoffrey Holmes Bernhard Pfahringer Peter Reutemann Ian H. Witten Mark Hall, Eibe Frank. The weka data mining software. SIGKDD Explorations, 11(1), 2009.


