DETECTING AND ANALYZING DDOS ATTACK USING MAP REDUCE IN HADOOP

G.S. NAVALE, VIVEK KASBEKAR, VIJAY GANJEPATIL, SHRAVANTI BUGADE

Abstract— An assault on a network that floods it with so many requests that regular traffic is either slowed or completely interrupted. Unlike a virus or worm, this can cause severe damage to databases. A Distributed Denial of service (DDoS) attack can employ hundreds or even thousands of computers that have been previously flooded by HTTP GET packet. The massive amounts of data that collect over time which difficult to analyze using common database management tools. Big data includes activity logs (machine generated data) which consist of unstructured format capture from web. The storage industry is continuously challenged as Big data increases exponentially where security is one of the challenging and harmful concern. To handle Big data Hadoop technology takes cardinal part in analysis. In this paper we have proposed detection of DDoS attack by using Counter based algorithm and Access Pattern algorithm which will implemented in Hadoop framework. Along this we can provide future prediction functionality using analytics. Dashboard provides visual view which will help to unveil the attacker and loyal user along with statistics.

Keywords- DDoS, HDFS, UDF, Algorithm, Analytics, Dashboard.

I. INTRODUCTION

A Denial of Service (DoS) attack is an attempt to make a computer resource unavailable to normal users. The DoS attacks are becoming more powerful due to bot behavior. Attack that leverages multiple sources to create the denial-of-service condition is known as The Distributed Denial of Service (DDoS) attack. DDoS attacks are big threats to internet services. HTTP flooding attack is one of the typical DDoS attack, in that hosts are sending large amount of request to target website to exhaust its resources[1]. Now a day there is massive growth in internet traffic. Due to this many DDoS attack detection systems facing a problem. DDoS attack detection systems are categorized into two parts software based system & hardware based system. According to paper proposed by Jinghe Jin & team they implemented the hardware based system for DDoS attack detection[2]. One of the major advantages of hardware-based DDoS defense system is that they can process packets at a higher speed but problem with this system are high false positive rate. In order to address these problem of false positive rate & big data traffic we are implementing software based system by which we can solve problem of high traffic rate. In this paper we are using hadoop framework in terms to address the problem of big data which is caused by DDoS attacks. We are proposing Counter based algorithm to detect DDoS attack to solve the problem of high false positive rate and Access pattern based algorithm to detect DDoS attack using behavior of attacker. We have proposed above mentioned DDoS detection algorithms along with future predictions i.e. predictive analysis using Revolution analytics (‘R’) and GUI will be implemented using dash-boarding tool since it supports organized business with meaningful & useful data.

II. HADOOP FRAMEWORK

Hadoop is an open-source software framework that supports data-intensive distributed applications. It enables applications to work with thousands of computationally independent computers and with petabytes of data[3], [4], Hadoop increases the storage space and the processing power by uniting many computers into one.

A small Hadoop cluster will include a single master and multiple worker nodes (slaves) as in Figure 1. The master node consists of a Job Tracker, Task Tracker, Name Node and Data Node. A slave or worker node acts as both a Data Node and Task Tracker. In a large cluster, HDFS is managed through a dedicated Name Node server to host the file system index and a secondary Name Node that can generate snapshots of

Figure 1. Hadoop Framework

Detecting And Analyzing Ddos Attack Using Map Reduce In Hadoop
the Name Node’s memory structures, thus preventing file system corruption and reducing loss of data.\(^5\)

A. Hadoop Distributed File System and Map Reduce

Hadoop Distributed File System (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Map Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

III. DDoS Attack Detection With Hadoop

A. Data Generation

Flume’s architecture is simple, robust, and flexible. In our system we are using flume tool for data generation or we can say log file generation. The below figure 2 shows a typical deployment of Flume that collects log data from a set of application servers. The deployment consists of a number of logical nodes, arranged into three tiers. The first tier is the agent tier. Agent nodes are typically installed on the machines that generate the logs and are your data’s initial point of contact with Flume. They forward data to the next tier of collector nodes, which aggregate the separate data flows and forward them to the final storage tier.

![Figure 2. Flume tool working](image)

B. Counter Based Algorithm

According to algorithms proposed by Yeonhee Lee and Youngseok Lee we are implementing this algorithms in hadoop framework.\(^6\)

Map Reduce algorithm to detect DDoS with URL counting. To lower the false positive rate, we adopted response rate against page requests. This algorithm needs three input parameters of time interval, threshold and unbalance ratio, which can be loaded through the configuration property or the distributed cache mechanism of Map Reduce. Time interval limits monitoring duration of the page request. Threshold indicates the permitted frequency of the page request to the server against the previous normal status, which determines whether the server should be alarmed or not. The unbalance ratio variable denotes the anomaly ratio of response per page request between a specific client and a server. This value is used for picking out attackers from the clients. In Map Reduce algorithm, the map function filters non-HTTP GET packets and generates key values of server IP address, masked timestamp, and client IP address. The masked timestamp with time interval is used for counting the number of requests from a specific client to the specific URL within the same time duration. The reduce function summarizes the number of URL requests, page requests, and server responses between a client and a server. Finally, the algorithm aggregates values per server. When total requests for a specific server exceeds the threshold, the Map Reduce job emits records whose response ratio against requests is greater than unbalance ratio, marking them as attackers. While this algorithm has the low computational complexity and could be easily converted to the Map Reduce implementation, it needs a prerequisite to know the threshold value from historical monitoring data in advance. We can use Java for implementation of counter based algorithm but it requires the thousand lines of code. So we can reduce this by using Pig Latin in which large java code is replaced by minimum lines of code. Pig provides a light-weight method for approximately measuring how much time is spent in different user-defined functions (UDFs) and Loaders. Approximate microseconds measures the approximate amount of time spent in a UDF and approximate invocations measures the approximate number of times the UDF was invoked.

IV. Detection DDoS Attacks With Analytics

Revolution Analytics addresses opportunities in Big Data Analytics while supporting the following objectives for working with Big Data Analytics:

1. Optimizing business process and reducing operational cost
2. Reducing the risk by anticipating and mitigating problems before they occur
3. Reduce data moment and replication along with optimize computational speed.

Traditional IT infrastructure is simply not able to meet the demands of new “Big Analytics” landscape. For these reasons, many enterprises are turning to the “R” statistical programming language and Hadoop (both open source projects) as a potential solution to this solve the commercial need. That’s why we are implementing Second Access pattern based algorithm in ‘R’.

A. Access Pattern Algorithm

To increase the efficiency of our system we are taking output of Counter based algorithm as an input to this algorithm. The access pattern-based detection method assumes that clients infected by the same bot conduct the same behavior and that attackers could be differentiated from normal clients the first job obtains...
access sequence to the web page between a client and a web server and calculates the spending time and the bytes count for each request of the URL; the second job hunts out infected hosts by comparing the access sequence and the spending time among clients trying to access the same server. Along with this we are using various algorithms in ‘R’ like regression algorithms, aggregation algorithms for predicting the future activity of detected attackers.

V. DASHBOARD

Dashboards give signs about a business letting you know something is wrong or something is right. The corporate world has tried for years to come up with a solution that would tell them if their business needed maintenance or if the temperature of their business was running above normal. Dashboards give you an overview of how your properties are performing by displaying summaries of different reports as widgets on a single or multiple pages through hyper link. It is difficult to collect, review, or analyze massive amounts of data can be overwhelming for anyone. That’s why industry leaders such as Microsoft and IBM devote so many resources to developing dashboard technology and dashboard software. We are here proposing a modern dashboard over more conventional data collection and visualizing methods such as manual recording or non-real-time manual input software. We are going to enactment the output of server which is in the format of log file using the modern dashboards technique provided by Tableau Public software. Tableau provide excellent feature over the traditional software such as it brings your data to life with interactive graphs, charts and maps that will engage your readers. With a few clicks you can embed your interactive graphs, dashboards, maps and tables anywhere and share with everyone. It provides tremendous graphical statistic information in fraction of second which will helpful to dish out genuine user and attacker.

CONCLUSION

In this paper, we propose the system which consists of implementation Counter based and Access pattern algorithm by using Map Reduce in Hadoop. With this we are using analytics to predict the future behaviour of attacker. The better user interface provided by means of Dashboard.

REFERENCES