International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

# A Survey on Big Data and DDoS Attack

Mr.Gowtham T. K & Prof. Rakesh V. S
[1]8th Semester Student, Cambridge institute of technology, Bangalore, Karnataka, India.
E-mail: goutham.15cs405@citech.edu.in
[2]Professor, Cambridge institute of technology, Bangalore, Karnataka, India.
E-mail: rakeshv.cse@citech.edu.in

***Abstract:***

*Big data is the term for any collection of datasets so large and complex that it becomes difficult to process using traditional data processing applications. Everything now a day's become electronic and connected to internet to transmit and store data. That is way rate of data flow over the internet has been increased. The volume, velocity and verity (3v's) of data flow has introduced the new term called Big Data. Later on 2v's (value and veracity) are added by IBM and Oracle in Big data's definition. The technologies used by bigdata applications to handle the massive data are Hadoop, Map Reduce, and Apache Hiver. Most of companies adopted and shifted their business to Bigdata (Hadoop), But still some companies like Govt. organization and other security firms are hesitating to shift on it due to its security logs.*

*This research work is about availability threats like Distributed Denial of Service (DDoS) attacks. Security in this world of digital computing plays a typical role. Since all the operations are automated and large volumes of data are being maintained in the servers. We observe that there are significant differences in the histogram under different scenarios, so that attack detection based on packet attribute analysis will be effective.*

***Keywords: Big Data, Hadoop, DDoS Attack***

## 1. Introduction

Distributed Denial-of-Service (DDoS) is one type of cyber-attacks in which the victim receives a large amount of attack packets coming from a large number of hosts. As a result, the victim will be overloaded and eventually it will be unable to perform any normal functions. Hadoop was not designed for the enterprise environment so had not included security in its design. It was considered for Hadoop deployment that the network is fully secured and trusted in which it is used but as in enterprise environment the threats are day to day evolved and nothing is secured over the internet. Due to Hadoop using in public network the security in Hadoop has become the need. The major security threats are Confidentiality, Integrity and Availability. Currently, any counter measures are done manually. When an attack is reported, offline traffic analysis will be carried out to identify the possible attacks. After identification, new access controls will be set up to filter the attack packets. To tackle the issue of response time, we propose a new method to deal with DDoS: automatic detection of attack traffic. If the network can detect attacks automatically, the response time may be shortened and damages may be reduced. To establish the feasibility of our approach, patterns of normal traffic data and attacking traffic data are obtained. Then the distributions of packet attributes in normal condition and attacking condition are obtained and compared to find out the deviation of attributes under attack from normal condition. If any anomaly is found, it may facilitate the identification of attack packet signature. This research only addresses the Availability threats of Hadoop. Apache Hadoop has already introduced some modules to handle with Failover issue like Zookeeper, Journal Node etc. In this research work these solutions of Hadoop have been tested along with its effectiveness against availability threats.

## 2. A Feature of Big Data

The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, and combat crime and so on".

As we go across the various kinds of literature the most generally talked characteristic of big data is 5V's the 3 V which are common to all are volume, variety, and velocity. The figure1.1 below shows 5 V in big data.

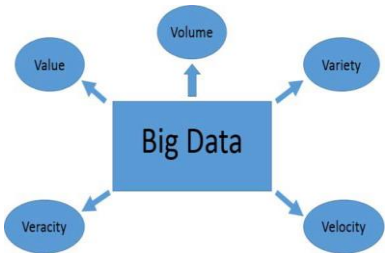



**Figure 1. FEATURE OF BIG DATA**

These 5 characteristics described as fallows

**Volume:** Volume refers to the incredible amounts of data generated each second from social media, cell phones, cars, credit cards, etc. The vast amounts of data have become so large in fact that we can no longer store and analyze data using traditional database technology.

**Velocity:** Velocity refers to the speed at which vast amounts of data are being generated, collected and analyzed. Every day the number of emails, twitter messages, photos, video clips, etc. increases at lighting speeds around the world. Every second of every day data is increasing.

**Value:** When we talk about value, we're referring to the worth of the data being extracted. Having endless amount of data is one thing, but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data.

**Variety:** Variety is defined as the different types of data we can now use. Data today looks very different than data from the past. We no longer just have structured data (name, phone number, address, etc.) that fits nice and neatly into a data table.

**Veracity:** Last, but certainly not least there is veracity. Veracity is the quality or trustworthiness of the data. Just how accurate is all this data? For example, think about all the Twitter posts with hash tags, abbreviations, typos, etc. and the reliability and accuracy of all that content.

## 3. Classification of Big Data

The characteristics of bigdata can be understood better by dividing it into classes description as fallows.

**Data Source:** The tremendously produced data is collected from the varied organization such as web and social sites, machines, sensing, transaction and IOT. For example in every 60 seconds 100,000+ tweets are made on twitter thus millions of people's data is generated every second.

**Content Format:** As the difficult type of data is being produced from various sources. The content formats are classified into structured, semi-structured and unstructured type. Structured is known which is basically in tabular format. Semi-structured data is one where the schema is not defined properly. Unstructured data is one who does not have any specific format like log file, video file, audio file.

**Data Stores:** These are the stores where the data generated from various sources is collected and stored. The data is stored in some categories like Document-oriented, Column-oriented, Graph based and Key-value.

**Data Staging:** this is the process in which there major steps is cleaning, normalization and transformation. Cleaning is unwanted data and useless data is being removed from the stored data. Further this cleaned data goes to normalization process it rearranged in the database. After the normalization process data is transformed into the human understand format.

**Data Processing:** The tremendously large and efficient data is being processed by using batch and real time data processing system. In the batch processing system the data is gathered processed and output is produced, but is this separate algorithm or program mandatory for input process and output. Where in real time data processing system there is no necessity for creating the distinct programs the continual processing of the input, process and output is done.

## 4. Hadoop

Hadoop is a framework that can run applications on systems with thousands of nodes and terabytes. It distributes the file among the nodes and allows to system continue work in case of a node failure. Hadoop makes it possible to run applications on systems with thousands of commodity hardware nodes, and the handle thousands of terabytes of data. Its distributed file system facilities rapid data transfer rates among nodes and allows the system to continue operating in case of node failure.

Hadoop was created by computer scientist Doug Cutting and Mike Carfella in 2006 to support distribution for the nutch search engine. It was inspired by google's Map Reduce, a software framework in which an application is broken down into numerous small parts.

As a software framework, Hadoop is composed of numerous functional modules. At a minimum Hadoop common as kernel to provide the framework's essential libraries. Other component include Hadoop Distributed File System (HDFS), which is capable of storing data across thousands of commodity server to achieve high bandwidth between nodes.

This protocol is an extension of Hadoop distributed file system. Hadoop is commonly used for distributed batch index building; it is desirable to optimize the index capability in near real time. Hadoop provides components for storage and

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

analysis for large scale processing. Now a day's Hadoop used by hundreds of companies. The advantage of Hadoop is Distributed storage & Computational capabilities, extremely scalable, optimized for high throughput, large block sizes, tolerant of software and hardware failure

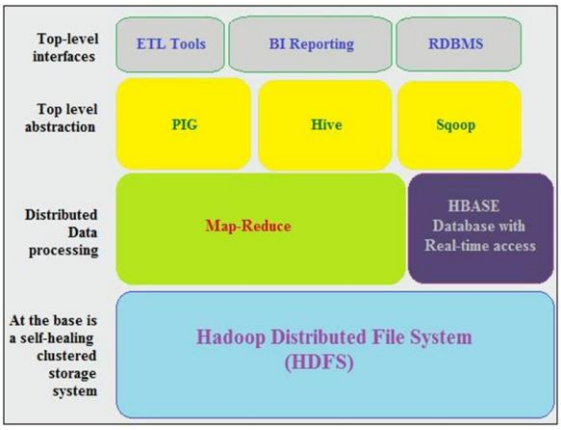



**Figure 2. ARCHITECTURE OF HADOOP**

## 5. Components of Hadoop

**HBase**: It is open source, distributed and Non-relational database system implemented in Java. It runs above the layer of HDFS. It can serve the input and output for the Map Reduce in well mannered structure.

**Oozie**: Oozie is a web-application that runs in ajava servlet. Oozie use the database to gather the information of Workflow which is a collection of actions. It manages the Hadoop jobs in a mannered way.

**Sqoop**: Sqoop is a command-line interface application that provides platform which is used for converting data from relational databases and Hadoop or vice versa.

**Avro**: It is a system that provides functionality ofdata serialization and service of data exchange. It is basically used in Apache Hadoop. These services can be used together as well as independently according the data records.

**Chukwa**: Chukwa is a framework that is used fordata collection and analysis to process and analyze the massive amount of logs. It is built on the upper layer of the HDFS and Map Reduce framework.

**Pig**: Pig is high-level platform where the MapReduce framework is created which is used with Hadoop platform. It is a high level data processing system where the data records are analyzed that occurs in high level language.

**Zookeeper**: It is a centralization based service that provides distributed synchronization and provides group services along with maintenance of the configuration information and records. **Hive**: It is application developed for data warehouse that provides the SQL interface as well as relational model. Hive infrastructure is built on the top layer of Hadoop that help in providing conclusion, and analysis for respective queries. Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant. It was originally developed to support distribution for the Nutch search engine project. Hadoop is open- source software that enables reliable, scalable, distributed computing on clusters of inexpensive servers. Hadoop is:

**Reliable**: The software is fault tolerant, it expects and handles hardware and software failures.

**Scalable**: Designed for massive scale of processors, memory, and local attached storage **Distributed**: Handles replication. Offers massively parallel programming model, Map Reduce.

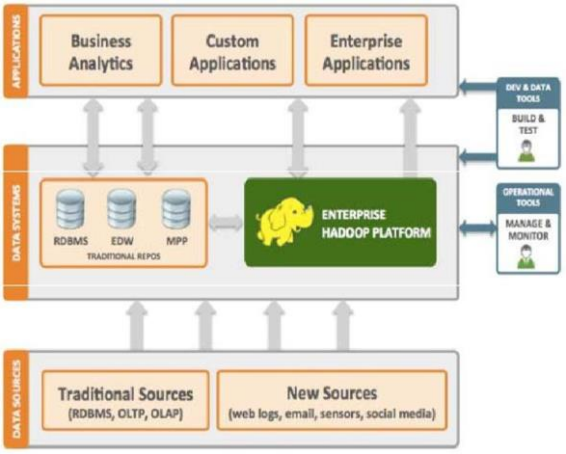



**Figure 3. HADOOP SYSTEM**

Hadoop is an Open Source implementation of a large-scale batch processing system. That use the Map-Reduce framework introduced by Google by leveraging the concept of map and reduce functions that well known used in Functional Programming. Although the Hadoop framework is written in Java, it allows developers to deploy custom-written programs coded in Java or any other language to process data in a parallel fashion across hundreds or thousands of commodity servers. It is optimized for contiguous read requests (streaming reads), where processing includes of scanning all the data. Depending on the complexity of the process and the volume of data, response time can vary from minutes to hours. While Hadoop can processes data fast, so its key advantage is its massive scalability.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018

## 6. DDoS History

In 20 years, distributed denial of service attacks have gone from a curiosity, to a nuisance, to a serious threat against the availability of websites, online services and applications. Today, anyone with a grievance and an internet connection can launch an attack. Easy to use tools and cheap attack services have made DDoS attacks more popular and more dangerous than ever.

Arbor Networks is now reporting that a US service provider suffered a 1.7Tbps attack earlier this month. In this case, there were no outages as the provider had taken adequate safeguards, but it's clear that the memcached attack is going to be a feature network managers are going to have to take seriously in the future.

DDoS attacks are initiated by a network of remotely controlled, well structured, and widely dispersed nodes called Zombies. The attacker launches the attack with the help of zombies. These zombies are called as secondary victims. The first massive DDoS attack has been encountered in the late june and early july, 1999 followed by an Fapi tool attack in 1998 which is not well documented. The first DDoS attack was to flood a single computer in University of Minnesota. The occurrences of DDoS attacks with year are given in the table 1. The servers suffered from DDoS attacks during the year 2000 [2] are Yahoo server, Amazon, Buy.com, CNN, and eBay, E*Trade and ZDNet, and NATO sites.

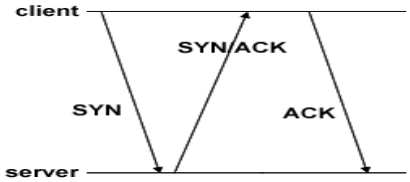
**Table 1. ORIGIN OF DDoS ATTACK**

| DDoS tool | Possible Attacks | Year |
|---|---|---|
| Fapi | UDP, TCP(SYN and ACK) and ICMP floods | 1998 |
| Trinoo | Distributed SYN DoS attack | 1999 |
| Tribe Flood Network | ICMP flood, SYN flood, UDP flood, and SMURF style attacks | 1999 |
| Stacheldraht | ICMP flood, SYN flood, UDP flood, and SMURF attacks | 1999 |
| Shaft | packet flooding attacks | 1999 |
| mstream | TCP ACK Flood attacks | 2000 |
| Trinity | UDP, fragment, SYN, RST, ACK and other flood attack | 2000 |
| Tribe Flood Network2K | UDP, TCP, and ICMP Teardrop and LAND attacks | 2000 |
| Ramen | Uses back chaining model for automatic propagation of attack | 2001 |
| Code Red & Code Red II | TCP SYN attacks | 2001 |
| Knight | SYN attacks, UDP Flood attacks | 2001 |
| Nimda | Attacks through email attachments, SMB networking and backdoors attacks | 2001 |
| SQL slammer | SQL code injection attack | 2003 |
| DDOSIM-0.2 | TCP based connection attacks | 2010 |
| Loris | Slowloris attack and its variants viz. Pyloris | 2009 |
| Qslowloris | Attacks the websites eg: IRC bots, Botnets | 2009 |
| L4D2 | Propagation attacks | 2009 |
| XerXeS | WikiLeaks attacks, QR code attacks | 2010 |
| Saladin | Webservers attacks, tweet attacks | 2011 |
| Apache killer | Apache server attacks, scripting attacks | 2011 |
| Tor' Hammer | HTTP POST attacks | 2011 |

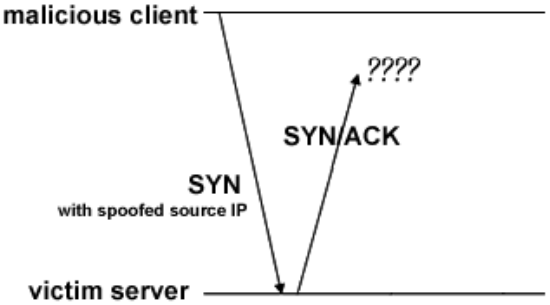
## 7. Anatomy of DDoS

In this section we describe the type of DDoS attack captured in the measured data from the ISP backbone network. The establishment of a TCP connection typically requires the exchange of three IP packets between two machines in an interchange known as the **TCP Three-Way Handshake.**

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 15
May 2018




In a traditional SYN Flood attack, a malicious client sent a SYN packet with a fraudulent source IP address. As a result, the SYN/ACK packet sent by the victim server will not get a reply as shown below.




In a DDoS TCP SYN Flood attack, the malicious client first infects a group of innocent clients called "zombies" and then launches a coordinated attack on the
victim.

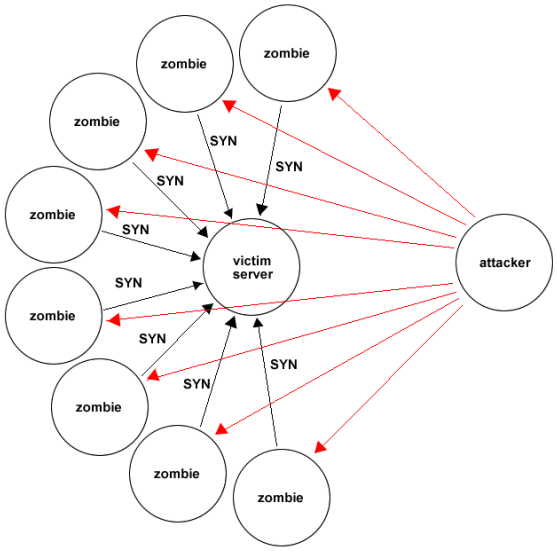



**Figure 4. ANATOMY OF DDoS**

## 8. Taxonomy of DDoS

Variety of DDoS attacks are sprouting in the computing world. The taxonomy of the DDoS attacks has been depicted in the Figure 5.

**1) Bandwidth Depletion Attacks**

This type of attack consumes the bandwidth of the victim by flooding the unwanted traffic to prevent the legitimate traffic from reaching the victim's network. Trinoo is one of the DDoS tools that cause the Bandwidth depletion attacks. These attacks can be further classified as:

**a) Flood Attacks:** This attack is launched by an attacker sending huge volume of traffic to the victim with the help of zombies that clogs up the victim's network bandwidth with IP traffic. The victim system undergoes a saturated network bandwidth and slows down rapidly preventing the legitimate traffic to access the network. This is instigated by UDP and ICMP packets.

An UDP flood attack is initiated by following steps:

- An attacker sends a large number of UDP packets to the victim's random or specified ports with the help of zombies.
- On receiving the packets, the victim looks the destination ports to identify the applications waiting on the port.
- When there is no application, it generates an ICMP packet with a message 'destination unreachable'.
- The return packets from the victim are sent to the spoofed address and not to the zombies.

**b) Amplification attacks:** The attacker sends a large number of packets to a broadcast IP address. In turn causes the systems in the broadcast address range to send a reply to the victim thereby resulting in a malicious traffic. This type of attack exploits the broadcast address feature found in most of the internetworking devices like routers. This kind of attack can be launched either by the attacker directly or with the help of zombies.

The *Smurf attack* is caused by following steps:

- Attacker sends packets to a network device that supports broadcast addressing technique e.g. Network amplifier.
- ICMP_ECHO_RESPONSE packets are sent by the network amplifier to all the systems in the broadcast IP address range. This packet implies the receiver to respond with an ICMP_ECHO_REPLY.
- An ICMP_ECHO_REPLY message from all the systems in the range reaches the victim.

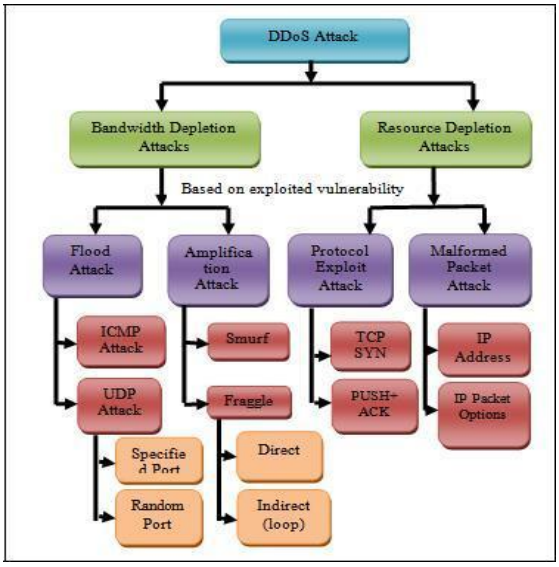



**Figure 5. TAXONOMY OF DDoS**

**2) Resource Depletion Attacks:**

The DDoS Resource depletion attack is targeted to strap the resources of the victim's system, so that the legitimate users are not serviced. The following are its types:

**a) Protocol Exploit Attacks:**These attacks is to consume the surplus quantity of resources from the victim by exploiting the specific feature of the protocol installed in the victim. TCP SYN attacks are the best example of this type.

**b) Malformed Packet Attacks:**The term malformed packet refers to the packet wrapped with malicious information or data. The attacker sends these packets to the victim to crash it. This can be performed in two ways:

- IP Address attack: The malformed packet is wrapped with same source and destination IP address thus creating chaos in the victim's OS. It rapidly slows down and crashes the victim.
- IP packet options attack: This attack makes use of the optional fields in the IP packet to form the malformed packet. The optional fields are filled by setting all the quality of service bits to one. So the victim spends additional time to process this packet. This attack is more vulnerable when attacked by more than one zombie.

## 9. Conclusion

The above study provides knowledge about the big data, its characteristics, features, and classifications. This paper possesses the basic and technical information about hadoop and its architecture. The big data handling techniques those handle a massive amount of data from different sources and improve overall performance of systems. The changes in the normalized frequencies of packet attributes due to attack is greater than the fluctuation of the normalized frequencies of packet attributes in normal condition. As DDoS attacks are on rise in all emerging technologies, we can expect a lot of security measures and corresponding vulnerabilities in future. This paper as a start provides a brief survey on DDoS attacks, taxonomy of attacks, its types and various counter measures to mitigate the DDoS attacks.

## 10. References

[1] Stephen M. Specht and Ruby B. Lee, "Distributed Denial of Service: Taxonomies of Attacks, Tools, and Countermeasures" Proceedings of the 17th International Conference on Parallel and Distributed Computing Systems, 2004 International Workshop on Security in Parallel and Distributed Systems, pp. 543-550, September 2004.

[2] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & McKinsey Global Institute. (2011). Big data: The next frontier for innovation,competition, and productivity.

[3] Upma Goyal, Gayatri Bhatti and Sandeep Mehmi, "A Dual Mechanism for defeating DDoS Attacks in Cloud Computing Model," International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 3, March 2013.

[4] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―Survey Paper On Big Data‖ International Journal of Computer Science and Information Technologies, Vol. 5 , 2014.

[5] Saraladevia, N. Pazhanirajaa, P. Victer Paula, M.S. Saleem Bashab, P. Dhavachelvanc Big Data and Hadoop - A Study in Security Perspective - International Symposium on Big Data and Cloud Computing (ISBCC), 2015 - DOI: 10.1016/j.procs.2015.04.091.