DOI: http://dx.doi.org/10.26483/ijarcs.v11i2.6507

Volume 11, No. 2, March-April 2020

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Internet Traffic Analysis: MapReduce based Traffic Flow Classification in Hadoop Environment

Sathish Kumar M
PG Student, Department of
Computer Science and Engineering
R.V. College of Engineering
Bangalore, Karnataka, India

Praveena T
Assistant Professor, Department of
Computer Science and Engineering
R.V. College of Engineering
Bangalore, Karnataka, India

Abstract: Internet is the global network that interconnects entities all over the world. This unparalleled network has occupied the mandatory part in the life of every individual. In recent days, due to the increase in the number of flow, the internet traffic is increased. The increasing traffic is flooding with the DDoS flows from multiple DDoS attackers. If DDoS flow traffic enters the internet, then there will be a drastic increase in the utilization of resources. Due to this, the legitimate traffic will not get proper service. In order to address the above issues, this paper has proposed an approach that classifies the internet traffic as Normal traffic flow or DDoS traffic flow. A huge volume of traffic flows is analyzed in this paper and the results are presented. The MapReduce is implemented for the classification as it accurately maps the flow features and reduces them into the appropriate traffic type. The incoming traffic is classified into one of the three categories as Web Traffic, DDoS Traffic (Heavy User) or DDoS Traffic (Spoofed IP). The main objective of this paper is to classify structured as well as unstructured data of IP, TCP, HTTP and NetFlow analysis. The experimental observations were carried out in the Hadoop 2.7.2 environment. The dataset is obtained from Wireshark, which consists of traffic flow based on latest traffic pattern. Hadoop Distributed File System (HDFS) and Map Reduce components of Hadoop are used under the metrics as Work Completion Time, Throughput and Accuracy.

Keywords: DDoS Traffic, Hadoop, Internet Traffic, MapReduce, Spoofed IP

I. INTRODUCTION

Internet traffic is becoming one of the serious threats in today's environment. With the increase in population, the number of people who use internet has also been increased to a great extent. Monitoring the internet traffic prove to be useful in evaluating the activity of the web page. Internet has become a mass communication that is being utilized worldwide [1]. Since, the internet traffic by default consists of huge volume of data, storage remains a question. The solution lies in the HDFS component of Hadoop. The Hadoop environment is open source software that encounters a lot of attacks these days [2]. Some of the various attacks include Botnets, Data leakage, DDoS and so on. There are many research works in order to solve these issues. The methods deployed include P2P network maps, special Hadoop component (Chukwa) to analyze the log, MapReduce cluster, implementation of Meta heuristic algorithms (Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), etc.). Hadoop Cluster deployment is necessary to understand the working of the system. The client can directly interact with the architectural model [3].

Of all the above mentioned attacks, Distributed Denial of Service (DDoS) is one of the most serious threats for any kind of networking. It is hard to predict and remove after occurrence. This type of attack occurs where the users request for service and receive response. It is observed that after the birth of internet DDoS has come into existence [4]. Most of the time, they have their target as servers of websites, banks, educational and governmental institutions. Certain researchers attempted to avoid DDoS using handshake mechanism and other such techniques. In practice, there is not an exact method to avoid the DDoS attack completely.

Nowadays, there is a great increase in the usage of internet services with the HTTP protocol. Many applications such as live streaming, music, games, and online interactions mostly prefer the HTTP protocol over TCP, UDP etc. As a result, attackers deploy the HTTP protocol to hack into the system. The string pattern matching of the URL is done to manually check the HTTP-based service [5], [6].In spite of normal approach, researchers started implementing algorithms to detect the presence of DDoS. Identifying the presence of malicious users will provide an alert to prepare the preventive measures. Gather the users in the network and their traffic log to monitor their activity and then detect the presence of abnormality [7]. Several classification algorithms like Principal Component Analysis (PCA), Support Vector Machine (SVM) and so on were utilized for classification purposes [8].

Later, in recent days, researchers moved on to the implementation of machine learning approaches. Several machine learning approaches are implemented for the classification purpose in the past few years. These classifications involved only two categories as normal and abnormal data. The popular long short-term memory (LSTM), Convolutional Neural Network (CNN), Deep Neural Network (DNN) and others have been used. Then, these methods are integrated with each other in order to detect the anomaly [9], [10], [11]. Even after introducing various approaches for DDoS detection and prevention, there are many unsolved issues that still remain as a major cause of internet traffic. The traffic is based on the n-number of visit to the server. Therefore, the count and interval traffic data is compared. The packets, flow and session are treated as significant core network and structural components. The modeling of internet traffic came into existence for the above mentioned components [12].

Internet traffic consists of both structured (type of data is predictable) as well as unstructured data (unarranged and unpredictable data). The process of analyzing and classifying the structured data is quite simple compared to the unstructured one. The amount of data being generated is uncontrollable. The location and time of congestion were also classified during the phase of analyzing the internet traffic. The data from social media sites like twitter were collected and classified [13]. Also, internet traffic reaches its peak when the traffic of social media is calculated as it is where people remain active 24*7.

In this paper, a classification system is proposed to predict the type of traffic. The main objective of the system is to achieve accuracy even when the number of flows increases. The major contributions are listed as follows:

- Implementing three different components as Traffic Collector & Loader, Reader and Traffic Analyzer for performing the process of the proposed system.
- Analyzing the Internet traffic and classifying them as Web traffic, DDoS traffic (Heavy User) and DDoS traffic (Spoofed IP)
- Implementation of Hadoop components as HDFS for storage and MapReduce for the purpose of classification of type of the flow present in the traffic
- Finally, experimental implementation and results in terms of Work completion time, Throughput and Accuracy.

The rest of the paper is organized as follows: Section II provides a detailed background study on the various methods used for internet traffic classification and analysis along with their drawbacks. Section III clarifies the working of the proposed system accompanied by the appropriate architecture. Section IV demonstrates the experimental setup, the implementation results and also a comparative study for the purpose of better understanding. Also, this section serves as the proof of proper functioning of the proposed system. Section V finish up the paper concepts along with drawing the future direction.

II. LITERATURE SURVEY

Article [14] proposed the comparison of DDoS attack detection in Hadoop environment based on two different configurations. In Hadoop single-mode configuration, a single node acts as both name and data node. When the DDoS attack affects, then the system has to be shut down completely and there was no recovery of the server. In Hadoop multimode configuration, the several networks are connected with the help of Network Interface Card (NIC). When the DDoS affects one network, the system will switch on to other network. This was because of the presence of more than one data node. As long as the Namenode was unaffected, the multimode system works fine. However, if more than one Datanode gets affected, the system cannot perform efficiently. The preventive measures from DDoS are very poor and didn’t deliver better security from the advanced attacks.

A research work have come up with the approach to detect the DDoS attack faster [15]. The work has been implemented in the Hadoop environment by considering the log file generated when a device uses the internet. These log files are stored in HDFS and verified for the presence of DDoS attack. The number of flows was counted and if it was greater than the threshold, then it was declared as DDoS. In spite of proper execution, this system suffers from the drawbacks of threshold value. Since, the threshold value was manually set and remains static, certain DDoS flows having values less than the threshold value will pass on through the system. In [16], Sufian Hameed et al. detected the presence of DDoS on a live traffic. The traffic was collected, transferred into log and then stored in HDFS. The MapReduce then executes the detection algorithm and identifies the DDoS attack. This work implemented clustering and parallel processing as well. Though the number of increased process ensures accuracy, the time consumption in setting up the system will be huge.

The article [17] attempted detect the DDoS in advance. The DDoS attack detection was performed for the four different attack profiles as constant rate attack, increasing rate attack, pulsating attack and subgroup attack. The classification of type of the traffic was based on Shannon metric. Since, the threshold was not fixed, various types of LR-DDoS and HR-DDoS attacks are detected. Here, the flows are not stored in HDFs rather assigned to mapper nodes directly and then the mapper collaborates with each other before sending results to the reducer. This system works fine when the number of flow was limited. If the number of flow was increases, the load on mapper node increases and the collaboration will not be accurate. In [18] explains that detection of DDoS in live traffic was faster than passive traffic, since the traffic collection phase consumed lot of time and resources. The various machine learning algorithms deployed in this paper include, K-Nearest Neighbor (KNN), Nave Bayes (NA), Random Forest (RF), Decision Tree (DT) and SVM. The algorithms are not as efficient as compared to the speed requirement of the approach. In order to detect a live traffic, these algorithms are too slow. The research in [19] detects and resolves the DDoS attack in Hadoop environment. The IP addresses of all the users are obtained and then, those IP belonging to the attackers are blocked. If the traffic contains more flows than the threshold, that particular IP address was marked as attacker. The normal users are alone allowed into the system. This method was more or less similar to [15] and suffers from the same drawbacks. Also, only the number of flows has been considered. The limited parameter consideration will not provide accuracy in detection.

Sughasiny in [20], utilized the in-memory processing technique to prevent the entry of malicious user into the network. Since Random Forest was a machine learning based classifier, the need for threshold has been replaced. But still, the classification accuracy was not a sure thing. Random Forest was nothing but collection of decision trees, handling tree structure was quite complicated when the number of flow was increased. In [21], the author proposed a cloud-based system to detect the attack in real-time in order to handle the traffic and protect the network. This paper holds traditional approaches and algorithms that were overcome in the recent years.

This paper work mainly focuses on classification of the type of the internet traffic. Even without the implementation of machine learning or Meta heuristic algorithm, the proposed system showed better performance and achieved best results. Thus, we provide a scalable and reliable system that ensures classification with high accuracy and acceptable work completion time.

III. PROPOSED SYSTEM

Internet traffic flow has affected the leading websites and lead to complete shutdown of those system. In this section, we have presented the working of the proposed system in detail. The proposed system provides the accurate classification in very short span of time for huge volume of traffic flow records. The main objective of the proposed system is to classify the traffic flow in the internet and avoid the congestion. In this work, we have considered the internet traffic flows present in the Wireshark dataset. We deployedthree different components for achieving the purpose of the system. The three components are: (a) Traffic Collector & Loader (b) Reader and (c) Traffic Analyzer.

A detailed architecture of the proposed system with various components and the work process is given in the Figure 1.

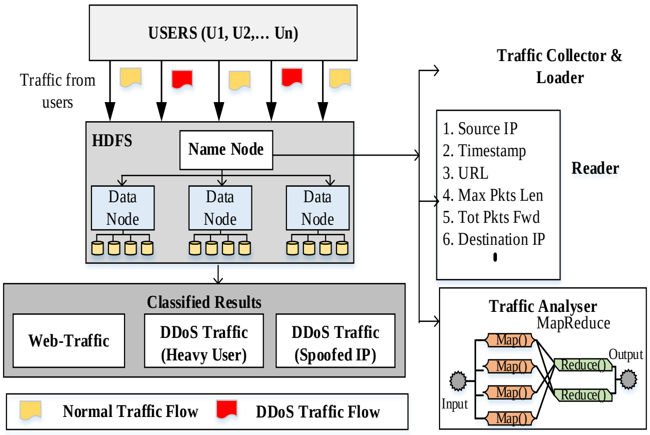

Figure1. Proposed architecture

The traffic collector is responsible for gathering the traffic flows until a particular period of time. The traffic loader will then load the flows dataset into the HDFS system. This is where the flow gets stored before being processed. The reader component will read the traffic flow and extract the features present in them. There are many packet features that provide better results for the classification. For the DDoS detection, the parameters used in the system are more than sufficient. Due to the limited number of parameters, the speed of execution of the proposed system is high. The reader will provide the flow features to the analyzer for proceeding further. Finally, the traffic analyzer will analyze the flow and classify them into one of the three categories. The three categories include Web traffic (T_1), DDoS Traffic (Heavy User) (T_2)and DDoS Traffic (Spoofed IP)(T_3).

Web traffic refers to the users request demanding service from the same website. In general, web traffic for a particular website is estimated based on the increased number of visits. But, in our work, we study the web traffic created by bots (attacker devices). Though the web traffic increases the tariff, it will greatly reduce the access of the website. The web traffic is identified by the URL having same web address. In order to classify this type of traffic, the source IP (src_IP), timestamp (tp) and Uniform Resource Locator (URL) are treated as the parameters.

In general, DDoS is an attack that occurs and affects the normal functioning of the network. The only motive of the attacker is to make the resource unavailable to the target user. DDoS issues arise as a result of frequent request from multiple DDoS attackers. As the name indicates, multiple attackers from different location will request for the same service frequently. The traffic consists of n-number of flows with increased length. DDoS Traffic (Heavy User), the Source IP (src_IP), Maximum Length ($Max\ Len$) of the Packet and the Total packets forwarded in the flow ($Tot\ Flow\ Frw$) are considered.

IP spoofing is done in order to impersonate another computer and thereby hide the identity of the actual sender. So that, even when the sender sends any unwanted content, the source IP points out to a different device. Spoofed IP mainly focuses on the obtaining valid impersonated IP of thelegitimate users in the network. And so, the IP address of the source and destination along with the timestamp is taken into account. Therefore, for DDoS Traffic (Spoofed IP), the Source IP (src_IP), timestamp (tp) and Destination IP (dst_IP) are calculated. With these features, the traffic flow is analyzed and all the flow gets classified. In this paper, we have given the pseudocode of the working of our proposed system. This provides a better understanding of the proposed work.

Pseudocode for Proposed Algorithm

```
Input: dataset (D)
Output: Classified traffic (T1,T2,T3,T4)
//    T4=Nrml_Traffic,
1. Begin
2. Load D from the traffic collector into HDFS
3. Read the flow F= {f1,f2,..., fn}// n represents the total number of flows in D
4. f = {src_IP ,tp ,URL , Max Len ,Tot Flow Frw , dst_IP }
5. Check f        // Start mapping the flows with the features in mapper
6. If f( srcIP ,tp ,URL = true)         {
Return T1 // Reduce function provides the output
Else
Go to Step 7
7. If f (srcIP ,Max Len ,Tot Flow Frw =true)        {
Return  T2 // Reduce function provides the output
Else
Go to Step 8      }
8.If f ( srcIP ,tp ,dstIP=true)       {
Return T3// Reduce function provides the output
Else
9. Return T4 // Reduce function provides the output     }
10. End
```

IV. EXPERIMENTAL ANALYSIS

A. *Implementation Environment*

In this paper, the proposed system is implemented using the Hadoop environment. Apache Hadoop 2.7 is open source software and is known for its service to the Big Data. This paper developed the system with respect to the setup involving the Ubuntu 14.04 LTS operating system environment along with the Java Development Kit 1.8, Netbeans 8.0 and Hadoop 2.7. The detailed system configuration is given in table-1.

Table I. Implementation Setup

HADOOP FEATURES		
Parameters	**Master/ Namenode**	**Slave/ Datanode**
Processor	Intel Core i3	Dual Core with 2.5GHz & above
CPU Core	4	2
RAM	4GB	2GB minimum
CPU Speed	3GHz	2.2 GHz
DATASET FEATURES		
Parameters	**Specification**	
Number of Flows	100	
Number of Features	89	

Based on the specification, the installation is carried out on the local system and then the experimental analysis is performed. The traffic flow is loaded from the Wireshark dataset whose specification is provided in Table-1. The flow pattern of the dataset matches with the recent traffic pattern as they are latest flow entries. Therefore, this system will classify the latest traffic pattern as well as the old one. Hence, the proposed system is flexible for any kind of data belonging to any time period.

B. Comparative Results

The experimental analysis is recorded in the form of graphical results and provided in this section. The experimental was conducted with a single Hadoop cluster consisting of single Namenode and Datanode. The results are presented as a comparison between the Web traffic, DDoS traffic (Heavy User), and DDoS traffic (Spoofed IP) in accordance with the Work completion time, Throughput and Accuracy. Work completion time denotes the time taken for predicting the type of the traffic for the incoming flow. Then, the throughput is measured as the sum of total number of flows transmitted within a specified time. All the three metrics are analyzed and then the flows are classified accordingly. The comparison is displayed in for each parameter in separate graphs for clarity.

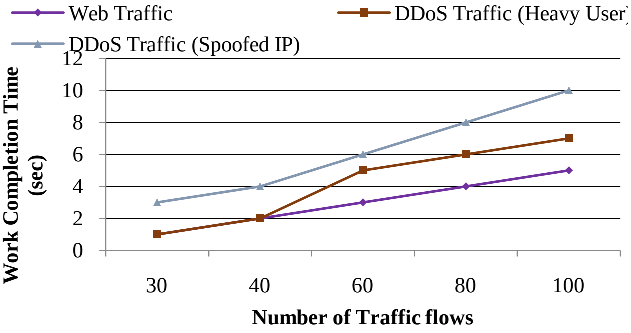

Figure 2. Comparison of Work Completion Time

Figure 2. Illustrates the work completion time of the classification based on the three different traffic types. It is noticeable from the figure that the web traffic is classified much faster than the other two categories and the DDoS traffic (Spoofed IP) is classified a little longer than the other two traffic types. This is because the presence of URL indicates it as web traffic whereas, in case of DDoS (Spoofed IP) the source and destination IP have to be analyzed and classified. From the graph, the web traffic of the hundred flows is classified in 3 seconds whereas the DDoS traffic (Spoofed IP) consumed 6.2 seconds respectively. This is only a linear raise and also guarantees that even when the number of flow increases, the work completion time will be classified more or less in the same range.

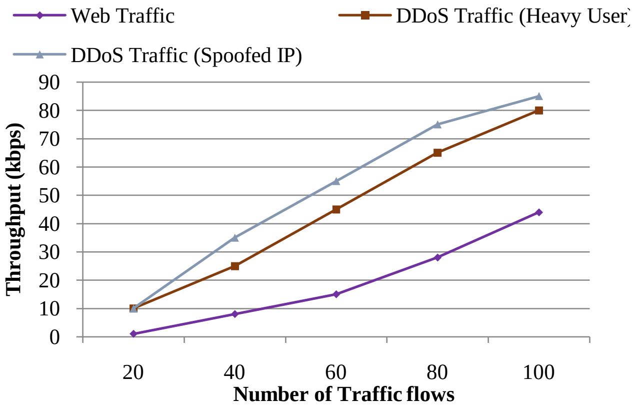

Figure 3. Comparison of Throughput

As discussed above, the throughput of the three traffic types is evaluated and compared in the Figure 3. The web traffic is transmits minimum number of packets than the other two categories and the DDoS traffic (Spoofed IP) is transmitting increased number of packets than the other two traffic types. This is because the nature of DDoS traffic is to transmit frequent traffic that consists of increased number of flows.

Also, the flows are transmitted with higher length than the length of the normal packets. Approximately 96 kbps is transmitted by the web traffic, 225 kbps is transmitted by DDoS traffic (Heavy User) and around 260 kbps is transmitted by the DDoS traffic (Spoofed IP). Hence, the web traffic transmits the flow much faster than the other two categories.

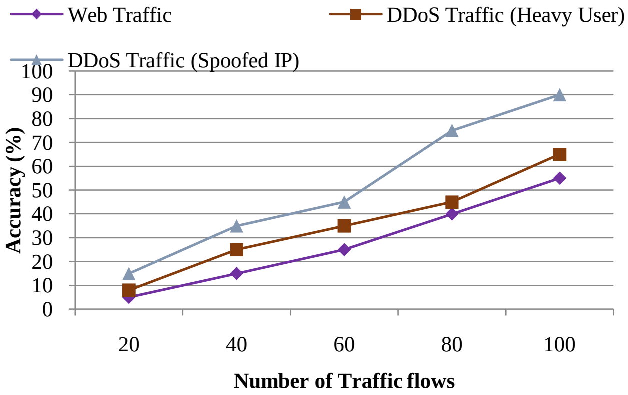

Figure 4. Comparison of Accuracy

Figure 4. Demonstrates the graphical analysis results of accuracy in the classification of all three traffic types. The correctness of the classification is denoted by the accuracy. It is viewed that the web traffic holds less accuracy than the other two categories and the DDoS traffic (Spoofed IP) holds increased accuracy than the other two traffic types. From the graph, it is observed that for a given 100 flows, the web traffic has been classified with 50% of accuracy, the DDoS traffic (Heavy User) flows has been classified with 60% of accuracy and the DDoS traffic (Spoofed IP) flows has been classified with 90% of accuracy. The accuracy here based on the improved performance of the MapReduce component. For any incoming traffic, the flow is mapped

accurately and reduced into its appropriate type. Therefore, even when the number of flow increases, the accuracy of this system will not go down.

Table II. Comparative analysis of Traffic Classification

Traffic types Parameters	Web Traffic	DDoS Traffic (Heavy User)	DDoS Traffic (Spoofed IP)
Work Completion Time	3 s	4.2 s	6.2 s
Throughput	19.2 kbps	45 kbps	52 kbps
Accuracy	28 %	35.6 %	52 %

An extensive comparison on the average analysis of the three metrics for the three types of traffic is given in Table 2. It is obvious from the table that the web traffic classification is completed within 3 seconds with an accuracy of 28% and transmits around 19.2 kbps of flows. Similarly, for the other type of traffic the same procedure is followed with the respective values. These values are given based on average for 100 flows. This table is a reassurance that even when the rate of flow present in the traffic rises drastically, there will not be a sudden increase in the values and at the same time the accuracy also will not go down.

V. Conclusion And Future Direction

The DDoS attacker will keep the server busy and disrupt the normal functioning of the network. Many DDoS attacks will halt the system with increased traffic flow. In this paper, a system is proposed to perform the classification of internet traffic. First of all, the traffic flows are collected and loaded into the system for storage and analysis. Then, the reader will read the traffic flows and obtain the features. These features are then given to the traffic analyzer. Here, the flows get classified as Web traffic, DDoS traffic (Heavy User) and DDoS traffic (Spoofed IP) respectively. In this paper, six significant flow features are considered for the purpose of DDoS classification. The final classification proved to be useful in identifying which are normal traffic flow and DDoS traffic flow. The experimental analysis was carried out in Hadoop environment and the results are obtained. The deployment of MapReduce ensured the accuracy of the proposed system. The appropriate mapping and reducing function lead to flawless classification of traffic flows. In addition, this paper also evaluated the hypothesis in terms of throughput as well as work completion time.

In future, the proposed work will be extended to detect the presence of other types of attacks as well. These attacks include various types of DDoS attack. Implementation of machine learning algorithm for the purpose of accurate detection is also planned. In addition, the work will be compared with the other existing works as well. This way, the worth of the proposed work can be proved.

VI. References

[1] Arthur Callado, Carlos Kamienski, Geza Szabo, Balazs Peter Gero, Judith Kelner, Stenio Fernandes, Djamel Sadok, “A Survey on Internet Traffic Identification”, IEEE Communications Surveys & Tutorials, IEEE, Vol. 11, no. 3, pp. no. 37-52, 2009.

[2] Akshay Kumar Suman, Dr. Manasi Gyanchandani, Priyank Jain, “A Survey on Miscellaneous Attacks in Hadoop Framework”, 2018 2nd International Conference on Inventive Systems and Control, IEEE, 2018.

[3] Ronaldo Celso Messias Correia, Gabriel Spadon, Pedro Henrique De Andrade Gomes, Danilo Medeiros Eler, Rogério Eduardo Garcia and Celso Olivete Junior, “Hadoop Cluster Deployment:A Methodological Approach”, information, MDPI, Vol. 9, no. 6, 2018.

[4] Kaushik Sekaran, G.Raja Vikram, B.V. Chowdar, UNP Gangadhar Raju, “Combating Distributed Denial of Service Attacks Using Load Balanced Hadoop Clustering in Cloud Computing Environment”, ICDTE 2018: Proceedings of the 2nd International Conference on Digital Technology in Education, pp. no. 77-81, 2018.

[5] Andrea Morichetta, Marco Mellia, “Clustering and evolutionary approach for longitudinal web traffic analysis”, Performance Evaluation, ELSEVIER, vol. 135, 2019.

[6] Neha Sehta, Karuna Mishra, “Network Traffic Classification Using Hadoop Server”, International Journal of Engineering Science and Computing, IJESC, Vol. 8, no. 10, 2018.

[7] Margaret Gratian, Darshan Bhansali, Michel Cukier, Josiah Dykstra, “Identifying Infected Users via Network Traffic”, Computers & Security, ELSEVIER, Vol. 80, pp. no. 306-316, 2019.

[8] Muhammad Aamir, Syed Mustafa Ali Zaidi, “Clustering based semi-supervised machine learning for DDoS attack classification”, Journal of King Saud University - Computer and Information Sciences, ELSEVIER, 2019.

[9] Tae-YoungKim and Sung-Bae Cho, “Web Traffic Anomaly Detection using C-LSTM Neural Networks”, Expert Systems With Applications, ELSEVIER, Vol. 106, pp. no. 66-76, 2018.

[10] Mohammed Ali Al-Garadi, Amr Mohamed, AbdullaAl-Ali, Xiaojiang Du, Mohsen Guizani, “A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security”, Cryptography and Security, arXiv, 2018.

[11] Alan Saied, Richard E. Overill, Tomasz Radzik, “Detection of known and unknown DDoS attacks using Artificial Neural Networks”, Neurocomputing, ELSEVIER, Vol. 172, pp. no. 385-393, 2016.

[12] Asad Arfeen, Krzysztof Pawlikowski, Don McNickle, Andreas Willig, “The role of the Weibull distribution in modelling traffic in Internet access and backbone core networks”, Journal of Network and Computer Applications, ELSEVIER, Vol. 141, pp. no. 1-22, 2019.

[13] Muhammad Taufiq Zulfikar, Suharjito, “Detection Traffic Congestion Based on Twitter Data using Machine Learning”, Procedia Computer Science, ELSEVIER, Vol. 157, pp. no. 118-124, 2019.

[14] ShakeelAhmad,AmanullahYasin, Qaisar Shafi, “DDoS Attacks Analysis in Bigdata (Hadoop) Environment”, 2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST), IEEE, 2018.

[15] Vishal Maheshwari, Ashutosh Bhatia and Kuldeep Kumar, “Faster Detection and Prediction of DDoS

attacks using MapReduce and Time Series Analysis", 2018 International Conference on Information Networking (ICOIN), IEEE, 2018.

[16] Sufian Hameed and Usman Ali, "HADEC: Hadoop-based live DDoSdetection framework", EURASIP Journal on Information Security, 2018.

[17] Nilesh Vishwasrao Patil, C.Rama Krishna, Krishan Kumar, SunnyBehal, "E-Had: A distributed and collaborative detection framework for early detection of DDoS attacks", Journal of King Saud University - Computer and Information Sciences, ELSEVIER, 2019.

[18] Awais Ahmed, Sufian Hameed, Muhammad Rafi, Qublai Khan Ali Mirza, "An Intelligent and Time-Efficient DDoS Identification Framework for Real-Time Enterprise Networks", Cryptography and Security, arXiv, 2020.

[19] Nakul Chorey, Rujuta Kate, Prajakta Khatavkar, Ms. Renuka.R.Kajale, "Detecting, Capturing &Resolving of DDoS Attacks with Hadoop", IJSRD -International Journal for Scientific Research & Development|, IJSRD, Vol. 6, no. 2, 2018.

[20] M. Sughasiny, "Zero Event Anomaly Detection in Big Data using Spark for Fast and Streaming Applications", International Journal of Pure and Applied Mathematics, Vol. 119, no. 15, 2018.

[21] Mounir Hafsa and Farah Jemili, "Comparative Study between Big Data AnalysisTechniques in Intrusion Detection", big data and cognitive computing, MDPI, Vol. 3, no. 1, 2018.