ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.10

Network Traffic Classification Using Hadoop Server

Neha Sehta[1], Karuna Mishra[2]
H.O.D. of I.T.[1], M.E. Student[2]
Department of Computer Science
Sushila Devi Bansal College of Technology, Indore, India

Abstract:
Network is the important part of any organization. Knowing the usage of computer network can help to manage network resources and internet traffic. Data mining algorithms can be used to analyze the trend of network traffic and observe the patterns. Major issue faced by existing work is large data generation. Processing on large data can't be an easy task and need advance technique to analysis purpose. Hadoop distribution server can be a good platform to analyze the large network data trends. In our work, analysis of network traffic using the Hadoop environment has been studied. A DBSCan based solution has been performed to observe the trend and load on individual system. The comparison made depends on single server and multi-node Hadoop cluster can for result evaluation purpose.

Keywords: DBSCAN, Data Mining, Hadoop, Network Traffic Classification.

I. INTRODUCTION

In current scenario, everyone is using the internet however, the latest tool for network examination, which handles the traffic of internet, is not growing with same pace. As we know that data on network is very vast and increasing every day in petabytes. With the time, it is realized that the traditional database is not enough for storage of large data. The needs and demands lead to innovation of bigdata. Big data is data sets, which consists of thousands of petabytes of data. Traditional tool does not process for large datasets. Hence, new innovative processors are also designed for accessing large amount of data.

As we know internet is derived name from network. Network performance is always a major concern of network administrator as well as network user. The traffic is increased in significant amount with the development of mobile, tabs at low prices. Every second user is accessing internet. In such large traffic the analysis is performed as well. This analysis of network is performed generally using single server, however when it comes to big data analysis single server is not just enough. The network classifier is the main component of network. It is used for analysis of packet as well as IP. In terms of memory, speed, processing capability. The classification of network is performed using payload signature. As payload consist of large set of information hence analyzing it consume large amount of data. In computer significant parameters are highly consumed.

DBScan algorithm is one of the widely used algorithms in data mining. In it clustering of similar data point is performed by clubbing them together by formation of cluster. The similar data points are connected in high-density point whereas dissimilar data is connected in low-density data. In dense region minimum points are kept and clustering is start from random point. If the points in high dense area increase then it is kept out from that section. However the most positive point of this algorithm is that it doesn't require number of cluster in advance like k mean.

Hadoop is open source highly reliable tool developed by Apache framework. The heart of Hadoop is MapReduce which is responsible for splitting of large size block and distribution to node in the form of cluster. In our work, we are analyzing the network traffic using the Hadoop environment. The comparison of using the single server and Hadoop cluster is then performed in order to reveal that our technique is better.

II. LITERATURE REVIEW

Zhao-wen LIN et al. [1] suggested the importance of Hadoop distributed system. In his work, the map reduce is explained which is the distribution of dataset and generation of output in such a way that the memory consumption should be less. Hadoop defines the distributed architecture importance as the processing enhances in considerable amount in it. When any unwanted source disturbed the slave node then other slave node which is disinfected perform the desired task, hence the process keep going.

Jiajia Chen1 et al. [2] designed the architecture in which analysis of web servers is given. As Internet is growing with fast rate the traditional technologies are not suitable to cop up with it. The number of users as well as devices is getting advanced in period. Our system suggests the technique in order to find the behavioral characteristics of servers and response time in which servers respond to request. The research is different than other research in the way that other research takes more time.

Jayeeta Dutta et al [3] redefine the importance of network analyzer. The paper suggests that for security purpose it is very much required to maintain the network classification. As machine learning is advanced field which work on analysis, but there work suggest that based on behavioral characteristics of the data the classification can be performed. The work is performed on Google Hangout, which is one of the very famous peer-to-peer applications.

TABLE 1
Comparative Study

Paper Title	Strength	Weakness	Applications
Analysis of Web Traffic Based on HTTP Protocol[1]	Provide mechanism to identify HTTP packets and classify the incoming and outgoing IP address.	It can't be considered as general solution. Can't classify the complete traffic	General level Internet traffic classification.
Application Traffic Classification in Hadoop Distributed Computing Environment[2]	They use classification mechanism to classify different types of packers.	Time based dataset. They don't consider the size of data which plays important role in large data processing.	Traffic classification and load estimation.
Network Traffic Classification in Encrypted Environment: A Case Study of Google Hangout[3]	Consider Hangout dataset as the primary source and attempt to classify hangout package from general network traffic.	Only concentrate on Hangout packet type	Can be expanded for other type traffic classification.
A Survey of Classification Algorithms for Network Traffic [4]	Provide details description of all type of classification techniques.	Do not provide any details about clustering or another mechanism.	Understanding of classification algorithms.
LASER: A Novel Hybrid Peer to Peer Network Traffic Classification Technique [5]	Uses Longest Common Subsequence (LCS)-based Application Signature Extraction technique, algorithm, a novel hybrid network traffic classification technique	Create too much overhead	It is one of the fines solutions for P2P network.
Profiling and Identifying Users' Activities With Network Traffic Analysis [6]	It converts users' network activities information into different sequences with frame size and interarrival time just like HMM model does.	The do not integrate any Hadoop concept for large data processing.	User data classification for network analysis.

III. PROBLEM DOMAIN

In large network, the analysis is quite tough. If the single server is used for analysis then analysis of network will be very cumbersome. The tough in the sense that network will require lot of memory, computation time and processing speed. The analysis of packet is required due to main reasons, the packet can have misguided information from the intruder, and the destination as well as routing information is also required. The network analyzer can be defined as the analyzer, which contains packet related information that how packet is travelled, which packet stops its retrieval in transit. For this analysis it requires flow as input and the output will be details of packets which are travelling in network. Hence the efficient system is required which gives the network analysis with less resource consumption.

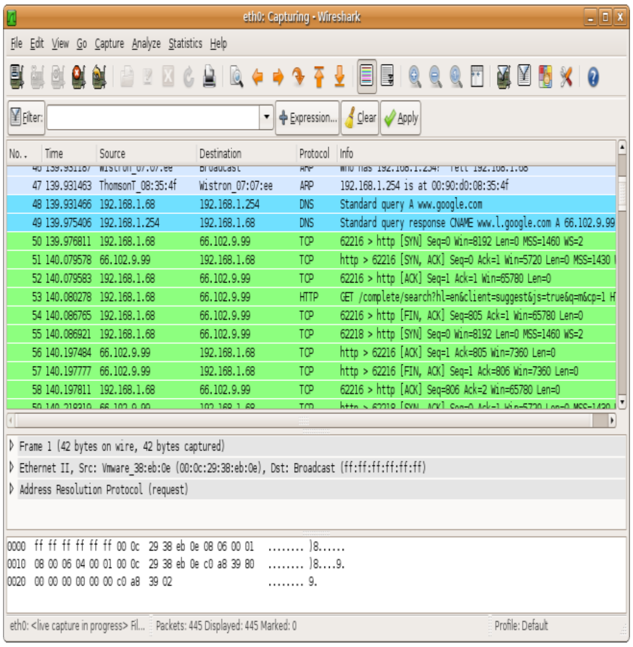
Study concludes that existing work faces challenge of large datasets and a solution is required to process those large datasets with the generation of data at great speed, because data production is increasing with higher percentage. If solution comes for this issue then it will provide higher performance of processing large data with maintaining security.

IV. SOLUTION DOMAIN

In our work, we are taking into account the college campus dataset. Entire traffic which travel in college will be considered in our work. As we know the college itself contain large amount of data hence the data is one of the large dataset. The entire flow is achieved using the tool named as Wireshark. The flow consists of details of sender and receiver like source IP address, source port number, destination IP address and its port number. Wireshark is open source tool, with graphical user interface with filtering option. Wireshark is used for capturing data and different networking protocol structure is observed. Some filter is used for displaying of data.

The workflow of our work is given as under:

- Initially the dataset is collected from SD Bansal College of Technology. Wireshark is the tool used for the retrieving all the datasets and resulting IP and packet transfer is shown.


eth0: Capturing - Wireshark
File Edit View Go Capture Analyze Statistics Help
Filter:
Expression...
Clear
Apply

```
No. .  Time        Source          Destination     Protocol  Info
    47 139.931463  ThomsonT_08:35:4f  Wistron_07:07:ee  ARP  192.168.1.254 is at 00:90:d0:08:35:4f
    48 139.931466  192.168.1.68    192.168.1.254   DNS       Standard query A www.google.com
    49 139.975406  192.168.1.254   192.168.1.68    DNS       Standard query response CNAME www.l.google.com A 66.102.9.99
    50 139.976811  192.168.1.68    66.102.9.99     TCP       62216 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
    51 140.079578  66.102.9.99     192.168.1.68    TCP       http > 62216 [SYN, ACK] Seq=0 Ack=1 Win=5720 Len=0 MSS=1430
    52 140.079583  192.168.1.68    66.102.9.99     TCP       62216 > http [ACK] Seq=1 Ack=1 Win=65780 Len=0
    53 140.080278  192.168.1.68    66.102.9.99     HTTP      GET /complete/search?hl=en&client=suggest&js=true&q=m&cp=1 H
    54 140.086765  192.168.1.68    66.102.9.99     TCP       62216 > http [FIN, ACK] Seq=805 Ack=1 Win=65780 Len=0
    55 140.086921  192.168.1.68    66.102.9.99     TCP       62218 > http [SYN] Seq=0 Win=8192 Len=0 MSS=1460 WS=2
    56 140.197484  66.102.9.99     192.168.1.68    TCP       http > 62216 [ACK] Seq=1 Ack=805 Win=7360 Len=0
    57 140.197777  66.102.9.99     192.168.1.68    TCP       http > 62216 [FIN, ACK] Seq=1 Ack=806 Win=7360 Len=0
    58 140.197811  192.168.1.68    66.102.9.99     TCP       62216 > http [ACK] Seq=806 Ack=2 Win=65780 Len=0

▷ Frame 1 (42 bytes on wire, 42 bytes captured)
▷ Ethernet II, Src: Vmware_38:eb:0e (00:0c:29:38:eb:0e), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
▷ Address Resolution Protocol (request)

0000  ff ff ff ff ff ff 00 0c  29 38 eb 0e 08 06 00 01   ........ )8......
0010  08 00 06 04 00 01 00 0c  29 38 eb 0e c0 a8 39 80   ........ )8....9.
0020  00 00 00 00 00 00 c0 a8  39 02                     ........ 9.

eth0: <live capture in progress> Fil...   Packets: 445 Displayed: 445 Marked: 0   Profile: Default
```

Figure 1: Wire shark scenario

- The data collected is raw data and can have variety of noise hence data need to be clean at initial. The data cleaning operation is performed on given data in order to remove inconsistent, incomplete data. Missing values may be present which is remove using this method.
- Once the data is cleaned then it is loaded to HDFS which Hadoop distributed file system. The HDFS is responsible for storing large data sets by streaming it at high bandwidth. User given task, here in this loading of college datasets is performed using HDFS. The HDFS in responsible for hosting as well as storing.

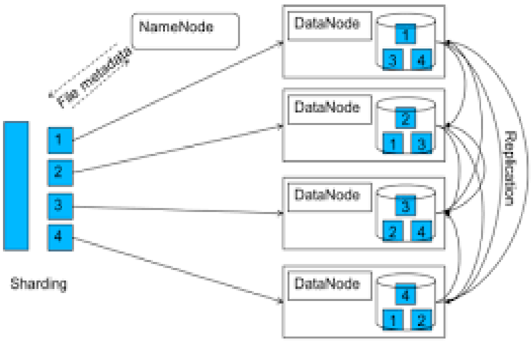


Figure 2: Hadoop Architecture

- Map reduce is function which takes dataset as an input as using the reducer return the reduced dataset as a result. In map reduce function the map node is loaded with very large memory whereas leaf node is loaded with less memory. In map reduce the target is to take the big data as input and reduce only desired data.
- Once the data is reduced, the filtering in given IP address is performed. For this the input given is the text file, initially the file analysis is performed in order to check whether the file is secured using any authentication scheme or not. Also flow is not continuous that flow is input after certain interval of time.
- The next important notification is that the flow is checked by observing first packet as first packet contains all the desired information. By taking all this as input the flow is generated using Field Extractor.

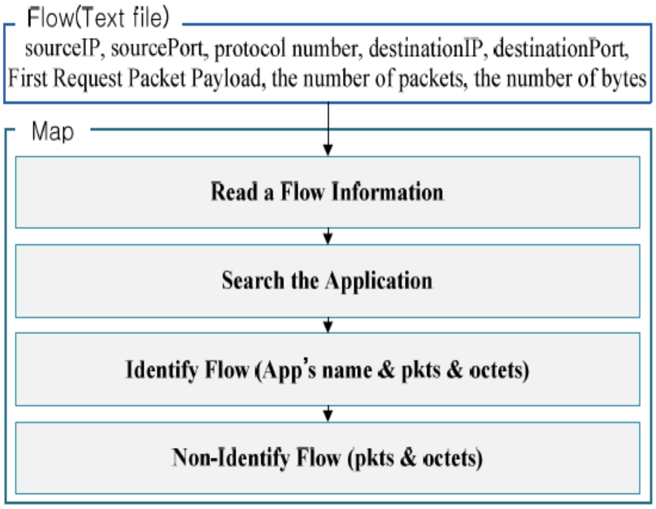

Flow(Text file)
sourceIP, sourcePort, protocol number, destinationIP, destinationPort,
First Request Packet Payload, the number of packets, the number of bytes
Map
Read a Flow Information
Search the Application
Identify Flow (App's name & pkts & octets)
Non-Identify Flow (pkts & octets)

Figure 3: System Workflow

V. IMPLEMENTATION STRATEGY

The process is performed using certain tables which are named as flow, packet, with type, total, classification, non-classification. The packet is given an input and the network in analyzed in certain interval. The application data consist of data packets and the observation is that with increase in packet size the processing time increased. Here we attaching the graph of it, and conclude that our architecture is better than single server approach. The approach is better in terms of memory and computation speed as well. Hence the Hadoop architecture is suitable as network analyzer. Following points are explored to draw the direction of implementation;

1. The complete study will be implementing in Java application along with Hadoop Server in Ubuntu 14.04 environment with one master and two slave nodes.
2. Four different size dataset in .csv format has been supplied as sample input and stored at HDFS for smooth data distribution.
3. Send MR and Receive MR fetch information from HDFS storage and process send and received packet information.
4. Afterwards, it forward information to Merge MR to combine the results and forward to DBSCAN MR.
5. DBSCAN MR prepares dynamic clusters and rewrite results on HDFS

VI. CONCLUSION

In our work, the network is analyzed using Wireshark tool. The key idea behind our concept is that the single server doesn't perform well when it comes to network analyzer. The reason for this is the network traffic is growing at such fast pace that single server cannot managed it. In our approach we are using Hadoop based architecture along with the Frame extractor tool in order to efficiently analyze network. The obtained result concludes that our system is better than existing system.

REFERENCES

[1] Jiajia Chen, Weiqing Cheng, "Analysis of Web Traffic Based on HTTP Protocol".

[2] Kyu-Seok Shim, Su-Kang Lee and Myung-Sup Kim, "Application Traffic Classification in Hadoop Distributed Computing Environment" published in Asia-Pacific Network Operation and Management Symposium (APNOMS) 2014.

[3] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers "Big data: The next frontier for innovation, competition, and productivity". McKinsey Global Institute. 15 May 2011.

[4] Shilpa and Manjit Kaur " BIG Data and Methodology-A review" International Journal of Advanced Research in Computer Science and Software Engineering.

[5] G P Sajeev, Lekshmi M Nair, "LASER: A Novel Hybrid Peer to Peer Network Traffic Classification Technique" published in Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India.

[6] Ma Tao and Ye Chun ming, Chen Juan, “Profiling and Identifying Users’ ActivitiesWith Network Traffic Analysis” Security Control Laboratory Jiaxing, China, 2015.

[7] Jayeeta Datta, Neha Kataria, Neminath Hubballi “ Network Traffic Classification in Encrypted Environment: A Case Study of Google Hangout ” Department of Computer Science and Engineering Indian Institute of Technology Indore.

[8] R.Deebalakshmi, Dr.V.L.Jyothi, “A Survey of Classification Algorithms for Network Traffic” published in Second International Conference on Science Technology Engineering and Management (ICONSTEM) 2016.

[9] Bing Liu, Machine Learning, Cybernetics,"A Fast Density-Based Clustering Algorithm for Large Databases",International Conference on , pp. 996-1000, 13-16 Aug.2006.

[10] Wu, Y. Jou, J. Zhang, X., “A Linear Dbscan Algorithm Based On Lsh”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007\

[11] LI Jian, YU Wei, YAN Bao-Ping,"Memory Effect in DBSCANAlgorithm”, In Proceedings of 2009 4th InternationalConference on Computer Science & Education, IEEE 2009.

[12] Md. Mostofa Ali Patwary, Diana Palsetia, Ankit Agarwal,Wei-keng Liao, Fredrik Manne, Alok Choudhary”, A New Scalable Parallel DBSCAN Algorithm Using the Disjoint-Set Data Structure”, In Proceeding SC '12 Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, IEEE 2012.