Internet Traffic Flow Analysis using Hadoop

Sudipa Biswas

Abstract

The internet traffic analysis elucidates the network administrator for monitoring the ongoing operation in the network and to understand the network so that the behavior could be examined and large problem can be examined. Flow analysis assists in traffic management, allocation of resources and fault tolerance. Due to the fast increase in internet user simultaneously the network usage has also escalated rapidly. The major problem of this fast growth in network is the traffic management, storing of traffic data and analysis this enormous amount of data in a single machine. To resolve this issue hadoop has been implemented to scan multiple input data and produce output for traffic identification and clustering flow. In this paper internet traffic flow analysis has been done using hadoop. In this proposed method system accepts packet data as input from network and this input is appended to hadoop distributed file system (HDFS) and at last processing is done through MapReduce. Once the output has been generated the network administrator analyses the internet traffic and troubleshoot any problem if necessary.

---

1 Introduction

nternet is inclusive system which connects numbers of computer with each other. It uses TCP/IP to get connect this devices which contains packets getting from source to destination computer. This computer network is usually administered by [1] software defined network (SDN).This assist the network in by decoupling that drive the outcome about the traffic is remitted. To implement the traffic analysis of the big data the collection needs to be done in order to measure the data for Varity sources. The huge amount of data which can be of any form like 3d data, audio, video, text and many more which cannot be processed by any traditional way but by the big data approach we can measure and analyze the data to be further analyzed for resolving network related issues. Hadoop [2] is implemented that uses basic programming to process large amount of data sets. The main intention of this paper is to design and implement a system to for network traffic analyze utilizing hadoop clusters [3]. Once the input is given the detailed measuring and analyze is done on the input and output is derived on the basis of the input given to the system.

2 II.

3 Proposed System a) Hadoop Framework

Hadoop is literally open source which is java based programming that will support analyzing processing and storage for large data set in a computing environment. Hadoop makes possible on the application on system with thousands of commodity hardware nodes and handling terabytes of data. Its file system which is distributed and facilities rapid data transfer rates among nodes and continue operating a nodes failure. Hadoop has emerged as big data foundation and scientific analytics, business and planning.

4 b) System Description

The process of flow analysis consists of mainly three main factors. Firstly the data exchange, secondly the analysis and thirdly the user interface. In the data exchange process the HDFS [4] is implemented to store the information or the data to be used as an input in analyzing. In the analysis process the data is analyzed and managed and other factors such as node, link, and flow analysis are also implemented in this process. In the user interface the user interface the system and graphical display is displayed to enhance the user understands the analysis flow. The API [5] and GUI [6] tool is also implemented which enhances the communication. The network gives the input to the system. Below in Fig1 component analysis flow architecture is given.

5 I

Abstract-The internet traffic analysis elucidates the network administrator for monitoring the ongoing operation in the network and to understand the network so that the behavior could be examined and large problem can be examined. Flow analysis assists in traffic management, allocation of resources and fault tolerance. Due to the fast increase in internet user simultaneously the network usage has also escalated rapidly. The major problem of this fast growth in network is the traffic management, storing of traffic data and analysis this enormous amount of data in a single machine. To resolve this issue hadoop has been implemented to scan multiple input data and produce output for traffic identification and clustering flow. In this paper internet traffic flow analysis has been done using hadoop. In this proposed method system accepts packet data as input from network and this input is appended to hadoop distributed file system (HDFS) and at last processing is done through MapReduce. Once the output has been generated the network administrator analyses the internet traffic and troubleshoot any problem if necessary.

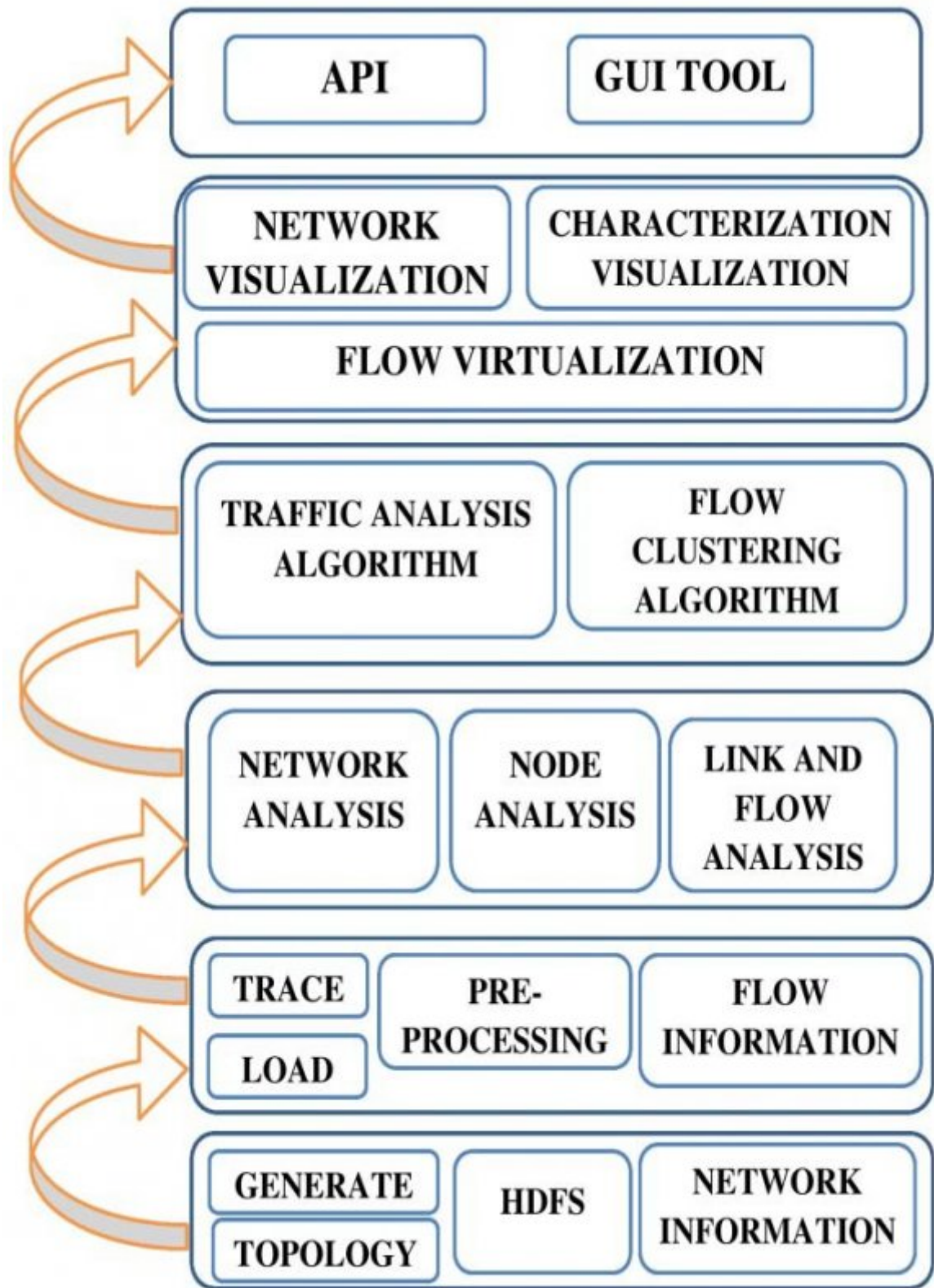# 6 Fig.1: Component analysis flow architecture

The input is actually data packet following from network. All the information is actually stored in HBASE. After that the system executes HDFs and MapReduce **??**7] function on the input data packet. After the function gets implemented and packets are stored the flow mechanism is implemented. The statistical view of the data can be viewed by the user. The different platform like Windows, Linux, and Mac OS/X are all on written in java platform. The specified format of the input helps in parsing of the input data. The sorting of the input file becomes necessary if the required format is done available. Parsing of the input is based on source IP address [8], types of packet, destination IP address, source port address and destination port address. The input which has same source port and destination port will form and stored together and input from the same source IP address and same destination IP address is clustered and stored as unstructured data in the database. In Fig2 block diagram of flow system analysis.

7 III. Conclusion

In this paper we have done the flow analysis and identification on hadoop platform. We have provided detailed analyses of the input data packet and classification of the type. This paper shows the methodology for analyzing packet file and statistical analysis of the original input packet and flow. The future work of this paper can be worked on the various problems like more networks makes more congestion troubleshooting the problem using the hadoop technology. [1] [2]

API GUI TOOL

NETWORK VISUALIZATION CHARACTERIZATION VISUALIZATION

FLOW VIRTUALIZATION

TRAFFIC ANALYSIS ALGORITHM FLOW CLUSTERING ALGORITHM

NETWORK ANALYSIS NODE ANALYSIS LINK AND FLOW ANALYSIS

TRACE PRE-PROCESSING FLOW INFORMATION

LOAD

GENERATE TOPOLOGY HDFS NETWORK INFORMATION

2

Figure 1: Fig. 2 :

7 III. CONCLUSION

78 [Chou] , Ta-Yuan Chou .

79 [Wun-Yuan and Huang] , Wun-Yuan , Huang .

80 [Mambretti] , Hui-Lan Lee; Te-Lung Liu; Joe Mambretti .

81 [Chen] , Yuanqi Chen .

82 [Zhou and Taneja] , Yi Zhou , ; Shubbhi Taneja .

83 [Qin] , Xiao Qin .

84 [Lai] , Yang Lai .

85 [Laptev] , Nikolay Laptev .

86 [Shi Zhong ()] 'An Efficient Data Mining Framework on Hadoop using Java Persistence API'. Zhi Shi Zhong .
87   *10th IEEE International Conference on Computer and Information Technology*, 2010.

88 [Huang] 'Design and implementation of HDFS data encryption scheme using ARIA algorithm on Hadoop'.
89   Jianzhong Huang . *aHDFS: An Erasure-Coded Data Archival System for Hadoop Clusters": IEEE*
90   *Transactions on Parallel and Distributed Systems. 4. Youngho Song; Young-Sung Shin; Miyoung Jang; Jae-*
91   *Woo Chang*, (2017 IEEE International Conference on Big Data and Smart Computing (BigComp))

92 [Hao Chen; Feiyeh ()] 'Heterogeneous Interconnection between SDN and Layer2 Networks Based on NSI'. Jim
93   Hao Chen; Feiyeh . *31st International Conference on Advanced Information Networking and Applications*
94   *Workshops (WAINA)*, 2017.

95 [Choudhary and Narooka] 'Hugepage & Swappiness functions for optimization of the search graph algorithm
96   using Hadoop framework'. Sunita Choudhary , ; Preeti Narooka . *2016 IEEE International Conference on*
97   *Computational Intelligence and Computing Research (ICCIC)*,

98 [Zeng and Zaniolo] *Very fast estimation for result and accuracy of big data analytics: The EARL system*, Kai
99   Zeng , ; Carlo Zaniolo . p. .