ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE
Your ulti-MATE Research Paper !!!

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE) Vol. 2, Special Issue 12, April 2016.

AN EFFICIENT APPROACH TO TRACK NETWORK PACKETS USING BIG DATA ANALYTICAL TECHNIQUE

K.Poongodi[1], P.Suganya[2], M.Vishalini[3], V.Soundarrajan[4]

[1]Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode Email: Poongodikks@gmail.com
[2]Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode Email: sugudpm@gmail.com
[3]Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode Email: vishalini519@gmail.com
[4]Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode Email: svsoundar5115@gmail.com

Abstract— An Efficient Approach to Track Network Packets using Bigdata Analytical Technique is used for identifying the number of websites and the number of users using those websites. The proposed system uses RStudio to store the big data and uses RLanguage to process the big data. Big Data offers an end-to-end portfolio to reduce costs, to gain a competitive advantage, and to increase the speed of innovation. The fast retrieving technique is used to retrieve the requested packets from the source-location and the currently running packets are also checked. A series of the event packets are reported to the base station. In a particular location many systems are accessed and from those accessed systems huge amount of packets are produced. These packets are stored in a source location in the form of PCAP (Packet capturing). Using a technique called packet tracer (Wireshark) the packets are retrieved. In this application retrieval of packets is based on the IP address of the source system and destination system. A technique called packet sniffing is used to identify the number of users using various websites and also it is used to identify the fake packets and. The unstructured data provided by the users are efficiently handled through map reduce algorithm. Finally visualized data and a detailed report will be generated based on generating various types of charts using R language. R language is used to identify unique IP address which is used to denote X and Y axis in the chart. Shiny package is used to view these charts and user controls. R language and Shiny package both are used in Rstudio to make an overall view. Rstudio is a used as a package to run R language.

Keywords: Big data**,** GUI, Rstudio, R Language, Shiny, Wireshark, pcap, IP Address, Packets, Packet Tracing

1. Introduction:

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making and better decisions can mean greater operational efficiency, cost reduction and reduced risk.

Analysis of data sets can find new correlations "to spot business trends, prevent diseases, and combat crime and so on." Scientists, business executives, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitation in e- science work, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research.

Big data is a popular, but poorly defined marketing buzzword. One way of looking at big data is that it represents the large and rapidly growing volume of information that is mostly untapped by existing analytical applications and data warehousing systems. Examples of this data include high-volume sensor data and social networking information from web sites such as Face Book and Twitter. Organizations are interested in capturing and analyzing this data because it can add significant value to the decision making process. Such processing, however, may involve complex workloads that push the boundaries of what is possible using traditional data warehousing and data management techniques and technologies.

2. Literature Review:

Netflow data collected by wireshark tool that are in .pcap format and then converted in text format.The original data set is in plain text format and each line represents a record of several fields separated by comma. After compute of packet count and size of packet in fixed interval of time we analysis large data sets, small data sets according the block size of hdfs. Default hdfs block size is 128MB [1]. The power of Big Data analytics for analysis of network flows to detect malicious behaviour of network using Apache Hadoop. Traditional technologies fail to provide the tools to support long-term, large-scale analysis of network traffic because these tools are not as flexible as Big Data Analytics tool for data formats and leveraged like Big Data analytics [2].

In a packet server the frame updates can only be performed at packet boundaries. It is to show that an update instant can be found in a packet server based only on the timestamps of the queued packets. Recall that the timestamp of a packet denotes the potential of the corresponding session at the instant that the packet completes its service in the fluid system [4].

In order to achieve more efficient packet scheduling, the following important factors for packet assignment are considered: the QoS such as delay and the required information bit rate [15].

3. Proposed System

In order to overcome the drawbacks of the existing system, the Intelligent system for Analysing and Eliciting Public Grievances is designed. Its primary purpose is to provide the efficient way to handle the data provided by the users. Due to more number of complaint data becomes big data which is difficult to store and process so HDFS is used to store the big data and uses MapReduce to process the big data.

The fast retrieving technique is used to retrieve the requested packets from the source- location and the currently running packets are also checked. A series of the event packets are reported to the base station. In a particular location many systems are accessed and from those accessed systems huge amount of packets are produced. These packets are stored in a source location and the required packets are also retrieved parallellely. Using a technique called packet tracer (WireShark) the packets are retrieved. In this application retrieval of packets is based on the IP address of the source system and destination system. A technique called packet sniffing is used to identify fake packets. The packets are sent through several paths and if any duplicate packets are identified, the unwanted packets are removed and the remaining packets will be stored in the base station. The unstructured data provided by the users are efficiently handled through RStudio. Finally the data are visualized and the detailed report will be generated based on the various types of charts.

The real benefit of report generation is that when the legislation plans of the government are manipulated, they are based on the report of Intelligent System. Any changes made on the complaint gets immediately reflected in the report. Finally, the fact is that the report outputs are presented as a Microsoft Excel Worksheets in the form of .CSV file and also in the form of charts using Shiny Package which makes it easier to distribute the information.

There are three main factors when selecting a tool for automated report generation from database. First, the cost of the tool and second the demands for simplicity of use and finally the flexibility of the generated reports and the features available.

Existing System

The existing scheme is made up of traditional database methodologies like SQL. It is capable of handling only structured data with the limited amount of data like gigabytes of data. In the existing system there is a need of dedicated centralized server with high configuration and the connected nodes are in same working environment with homogenous configuration. The analytics processes are performed on the preconfigured area. Users or clients should be present at that time in the environment. Networking concepts were used in the existing system for storing the data which is not efficient. Final outputs were not exact but they were produced based on assumptions.

The output produced was not user friendly. The report produced was only an information.

Network Analysis

Network Analysis is the process of intercepting and examining messages. The packets that are sent from source location to destination location are analyzed using this network analysis process. This is done using both wireless and wired. We can analyse the path through which the packets are sent. Using this process we can easily identify the fake packets.

Wireshark

It is an open source and is used as a packet analyzer. It is used to retrieve the packets from the network. It is used for troubleshooting networks, analysis, software and communication protocol development and also for education purposes. Wireshark is a program that understands the structure of different networking protocols. It can parse and display the fields along with the specified network. Wireshark uses Pcap to capture packets, so it can capture only packets on similar types. Data can be captured from the wired and wireless network. Captured network data can be browsed via GUI, using TShark. It differentiates different types of networking protocols by varying colors.

Packet Capture(pcap)

Pcap means Packet Capturing. This Pcap is used in Wireshark. Pcap is used convert the packets into .CSV file using TShark. Using Pcap we can retrieve the particular fields that we require. Pcap uses libpcap as a base to run. Incase if the Pcap runs in windows, Winpcap supporting file is used to capture the files in the form of Pcap. Both Libpcap and Winpcap files are same, so if one is there other is not needed. It is an Application Protocol Interface(API) which is written in C. If this needs to be implemented in other languages it requires wrapper.

Ubuntu

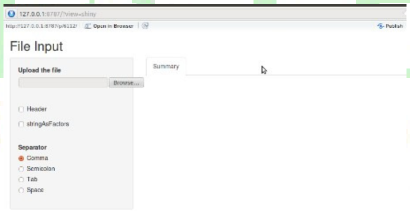
Ubuntu is a Debian-based Linux operating system and used in distribution for personal computers, smartphones and network servers. It uses Unity as its default desktop environment. It is an open source operating system and freeware. It named after the Southern African philosophy of Ubuntu. Ubuntu is built on Debian's architecture and infrastructure, to provide Linux server, desktop, phone, tablet and TV operating systems

RStudio

RStudio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics. RStudio is written in the C++ programming language and uses the Qt framework for its graphical user interface.

R-Shiny Package

Shiny is a web application framework for R language. It is a package that turn our analyses into interactive web applications. For this we don't need any programming languages like HTML, CSS or JavaScript. It is a package from RStudio that makes it incredibly easy to build interactive web applications with R. The features oh Shiny are: This has an attractive default UI theme based on Bootstrap. It uses a reactive programming model that eliminates messy event handling code, so we can focus on the code that really matters.



File Input

Upload the file

□ Header

□ stringAsFactors

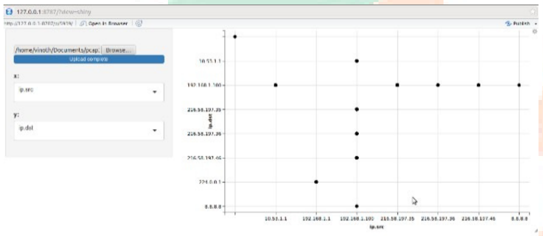
Separator

● Comma
○ Semicolon
○ Tab
○ Space

Fig. 1 File Upload Page in R Shiny application

Data Picturing

Data picturing is the demonstration of data in a graphical format. It enables decision makers to see analytics presented visually, so they can understand difficult concepts or identify different patterns. With interactive visualization, anyone can take the concept a step additional by using technology to compute problems into charts and graphs for detail understanding and interactively changing what data you see and how it's processed. It helps people to know the significance of data and its importance, also enables different dimensions of thinking and computing in the future for betterment of the result.


127.0.0.1:8787/?view=shiny

/home/vinoth/Documents/pcap

x: ip.src

y: ip.dst

Fig. 2 Visualized Data set in R Shiny application

.

4. Conclusion

The main motive of the intelligent system is to provide public satisfaction, meet the people needs and provide more sophistication. The Intelligent system serves its main purpose to get required packet from the people and to provide assurance in processing. The enhancements in the intelligent system include storing of data in RStudio platform. Unlike the existing system, the intelligent system uses RLanguage to provide the efficient handing of packets.

For packet analysis, the designed representative statistics of computation modules are taken using RLanguage. From the experiments, it is shown that our tool outperforms a typical traffic analysis method running on a single server.

References:

[1]Amreesh kumar patel, D.S. Bhilare, Sushil buriya, Satyendra singh yadav(2015)," Big Data Analytics for Net Flow Analysis in Distributed Environment using Hadoop", International Journal of Research in Computer and Communication Technology, Vol 4, Issue 7

[2]C. Dobrea, F. Xhafa (2013)," Intelligent services for Big Data science", ScienceDirect on Future Generation Computer Systems, pp. 0167-73

[3]Chia-WeiLee, Kuang-YuHsieh, Sun YuanHsieha, Hung-ChangHsiao (2014), "A Dynamic Data Placement Strategy for Hadoop in Heterogeneous Environments", IEEE transaction on Big Data Research, Vol 1, pp 14-22

[4]`Dimitrios Stiliadis, *Member, IEEE*, and Anujan Varma, *Member, IEEE(1998),"* Efficient Fair Queueing Algorithms for Packet-Switched Networks*",* ieee/acm transactions on networking, vol. 6, no. 2

[5] Firat Tekinerl and John A. Keane (2013), "Big Data Framework", IEEE International Conference on Systems, pp.1494-1498

[6] Gueyoung Jung, Nathan Gnanasambandam (2012) "Synchronous Parallel Processing of Big data Analytics services to Optimize Performance in Federated Clouds" IEEE Transactions on Cloud Computing, pp. 263-268

[7] Haibo Zhang, Fredrik Österlind, Pablo Soldati, Thiemo Voigt, and Mikael Johansson, *Member, IEEE(2015),"* Time-Optimal Convergecast With Separated Packet Copying: Scheduling Policies and Performance*",* ieee transactions on vehicular technology, vol. 64, no. 2

[8] Hoang-VuNguyen, EmmanuelMüller, KlemensBöhm(2014)," A Near-Linear Time Subspace Search Scheme for Unsupervised Selection of Correlated Features", ", ieee transactions on Big Data Research,Vol.1,pp. 37–51

[9] Huijun Gao, *Member, IEEE*, and Tongwen Chen, *Fellow(2008),"* Network-Based_∞ Output Tracking Control*",* ieee transactions on automatic control, vol. 53, no. 3

[10] J. Shi Et Al (2013), "Scalable Community Detection in Masssive Social networks Using Map reduce". IEEE Transactions on Cloud Computing J. Res. & Dev. Vol. 57 No. ¾ Paper 12

[11] Jyh-Ting Lai, An-Yeu Wu, *Member, IEEE*, and Wen-Chiang Chen(2007)," A Systematic Design Approach to the Band-Tracking Packet Detector in OFDM-Based Ultra wideband Systems", ieee transactions on vehicular technology, vol. 56, no. 6

[12] Lang Hong, *Senior Member, IEEE(1999),* "Multirate Interacting Multiple Model Filtering for Target Tracking Using Multirate Models", ieee transactions on automatic control, vol. 44, no. 7

[13] Linquan Zhang, Chuan Wu, Zongpeng Li, Chuanxiong Guo, Minghua Chen, and Francis C.M. Lau (2013), "Moving Big data to cloud:An online Cost Minimizing Approach", IEEE Transactions On Selected Areas In Communications, Vol.31, No.12, pp.25-45

[14] Rupam, Atul Verma, Ankita Singh," An Approach to Detect Packets Using Packet Sniffing(2013)", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.4, No.3

[15] Sadayuki Abeta, Hiroyuki Atarashi, and Mamoru Sawahashi(2002),"Broadband Packet Wireless Access Incorporating High-Speed IP Packet Transmission" ,pp.842-848

[16] Song Gao, Linna Li , Wenwen Li , Krzysztof Janowicz , Yue Zhang (2014), "Constructing Gazetteers From Volunteered Big Geo-data based on Hadoop", Elsevier on Computers,Environments and Urban Systems,pp.1-15

[17] Timothy J. Bellerby(2013)," Searchlight: Precipitation Advection Tracking Using Multiplatform Low-Earth-Orbiting Satellite Data", ieee transactions on geoscience and remote sensing, vol. 51, no. 4

[18] Wei Tan, Ke Xu, *Senior Member, IEEE*, and Dan Wang, *Senior Member, IEEE(2014),* An Anti-Tracking Source-Location Privacy Protection Protocol in WSNs Based on Path Extension, ieee internet of things journal, vol. 1, no. 5

[19] Xin Luna Dong, Divesh Srivastava(2013)," Big Data Integration", IEEE Conference,pp.1245-1248