Hadoop based Network Traffic Analysis for Client Reputation Score Calculation

Hasmukh B. Domadiya[1] & Dr. Girish C. Bhimani[2]
[1]Assistant Professor, National Computer College, Jamnagar.
[2]Head, Department of Statistics, Saurashtra University, Rajkot.

Received: August 26, 2018 Accepted: October 10, 2018

ABSTRACT *In today's fastest growing Internet, millions of web resources are maintained and accessed on regular basis. At the same time, a huge amount of web resources are added to the Internet every day. In the world of digitization and computerization, the users are increasing rapidly leading to an extremely large number of users accessing the Internet. In such scenario, it is indeed for an organization, ISP or even a country to analyze the access patterns of users to determine users at risk or at wrong direction. We have proposed a Hadoop based network analysis system which will rank its users a reputation score. We selected Hadoop to make our system efficient and scalable to the extremely large volume of data. We further thought of producing a distributed solution for real life implementation. The reputation score is like a percentage which will summarize how properly a user is accessing the Internet. This paper discusses this proposed system.*

Keywords: *Network Traffic, Hadoop, Internet, Reputation Score*

1. INTRODUCTION

The usage of the Internet is growing rapidly. From the industries to the home users, everyday, billions of users access the Internet for official or personal purpose. Here comes the need of analysis. The Internet generates a large volume of additional information called usage logs. Such logs are maintained by ISPs, Network Administrators or individual users with the help of firewalls or log servers. The purpose is to back track the events such as violation of security, to count usage for the generation of bills or to find current trends in terms of interests of the users. For all these purposes it is required to analyze what web resources users have accessed [1].

Most of the firewalls and log servers carry individual analyzer module to count number of hits web resource wise and to list various other necessary events in the form of reports. These solutions are suitable for the cases where the volume of daily data is limited. In case of high volume data such as when the average usage is in TBs and number of users are in Lacks, it is inefficient to use such devices. As Hadoop is the ultimate solution for processing large volume of data, we tried to design a client reputation score calculation module which we analyze the user visited web resources and rank them in the score of 0 to 100. It is obvious that the more a user scores, more reliable, genuine and careful he is. Those with very low score can be considered as users at risks with intentional or unintentional malicious activities, violation of security, content violation etc. such users need immediately attention and action from the concerned authority as a proactive efforts towards securing the network from which he is using the Internet.

2. HADOOP FOR ISPS

2.1 Why Hadoop?

As per the statistics done by statista the number of Internet users in INDIA is growing extremely faster. The same situation is in rest of the world. Figure 1[2] shows the past, present and future users statistics for INDIA given by statista. One another statistics regarding digital population in INDIA is given by statista in Figure 2[3]. From both of these analysis we can easily deduct that the Internet users are growing rapidly and so their Internet usage. Such situation leads to generation of extremely large volume of Internet logs [2,3].

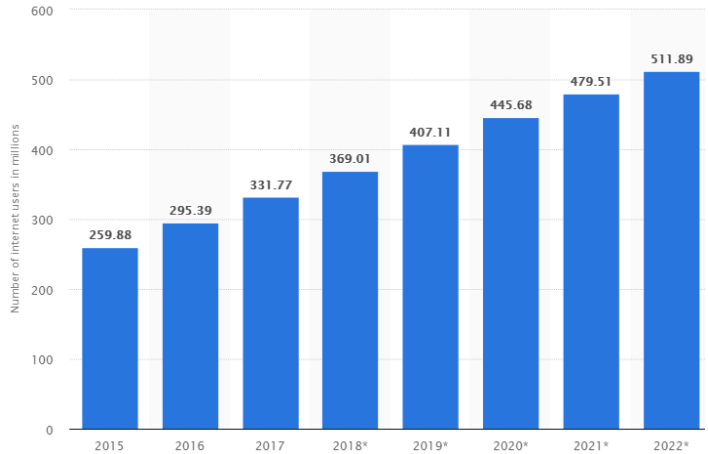

Figure 1 – Internet Users in INDIA[2]

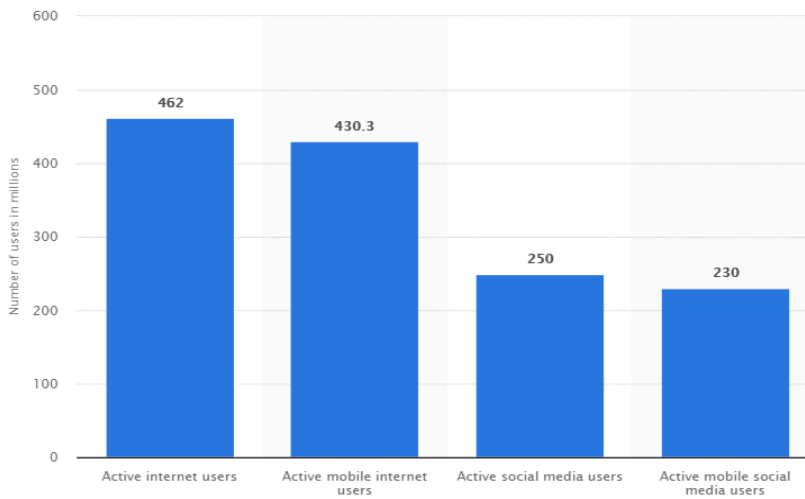

Figure 2 – Digital Population in INDIA[3]

Cisco Visual Networking Index: Forecast and Methodology, 2016–2021 report has predicted the internet traffic is going to be increased at extremely fast way. Some of the statistics shown in this prediction reports are shown in Table – 1[4] and Table – 2[4]. To process such huge data, Hadoop is best as per today's options [5].

Consumer Web, Email, and Data Traffic, 2016–2021	2016	2017	2018	2019	2020	2021	CAGR 2016–2021
By Network (PB per Month)							
Fixed web and data	6,795	7,467	8,569	9,610	10,706	11,337	11%
Mobile web and data	2,263	3,214	4,295	5,509	6,796	8,201	29%
By Geography (PB per Month)							
Asia Pacific	3,393	4,102	5,072	6,160	7,398	8,453	20%
North America	2,578	2,863	3,149	3,410	3,631	3,792	8%
Central and Eastern Europe	1,302	1,404	1,598	1,790	1,994	2,095	10%
Western Europe	693	901	1,177	1,450	1,692	1,882	22%
Middle East and Africa	469	732	1,038	1,358	1,728	2,189	36%
Latin America	624	680	831	953	1,059	1,128	13%
Total (PB per Month)							
Consumer web, email, and data	9,059	10,681	12,864	15,120	17,502	19,538	17%

Source: Cisco VNI, 2017

Table - 1 Global consumer web, email, and data traffic, 2016–2021[4]

Consumer Internet Video 2016–2021	2016	2017	2018	2019	2020	2021	CAGR 2016–2021
By Network (PB per Month)							
Fixed	38,369	51,022	65,413	83,172	103,341	125,988	27%
Mobile	3,660	6,094	9,696	15,010	22,512	33,173	55%
By Category (PB per Month)							
Video	29,325	39,518	51,722	68,279	89,181	116,905	32%
Internet video to TV	12,704	17,598	23,387	29,903	36,672	42,255	27%
By Geography (PB per Month)							
Asia Pacific	13,845	19,228	25,854	35,024	46,423	61,352	35%
North America	15,254	20,114	25,778	32,329	39,275	45,485	24%
Western Europe	6,290	8,520	11,005	14,035	17,533	21,760	28%
Middle East and Africa	1,170	1,944	3,068	4,754	7,218	10,895	56%
Central and Eastern Europe	2,527	3,350	4,369	5,824	7,754	10,170	32%
Latin America	2,943	3,960	5,035	6,215	7,650	9,500	26%
Total (PB per Month)							
Consumer Internet video	42,029	57,116	75,109	98,182	125,853	159,161	31%

Source: Cisco VNI, 2017

Table - 2 Global consumer internet video, 2016–2021[4]

2.2 Hadoop Cluster for ISP

Hadoop is the present ultimate solution to process extremely large volume of data accurately and in a distributed way. Figure 3 shows one such arrangement with reference of the requirement of network data analytics. Here we represent an actual scenario where an ISP has a set of offices zone wise. Here zone represents a state or a part of state serving users of a specific geographical area. Every zone has a slave node to keep logs of zone user activities. So if an ISP has n zones, corresponding n offices should be equipped with atleast total n hadoop slave nodes for data analytics. A central office could be equipped with a hadoop master node to coordinate and communicate with all the slave nodes. All these arrangement are backed by the network which is the Internet. Such arrangement can be considered as a single hadoop cluster to analyze network traffic. This arrangement is a general arrangement without being specific to what kind of network analytics we want to perform [5,6,7].

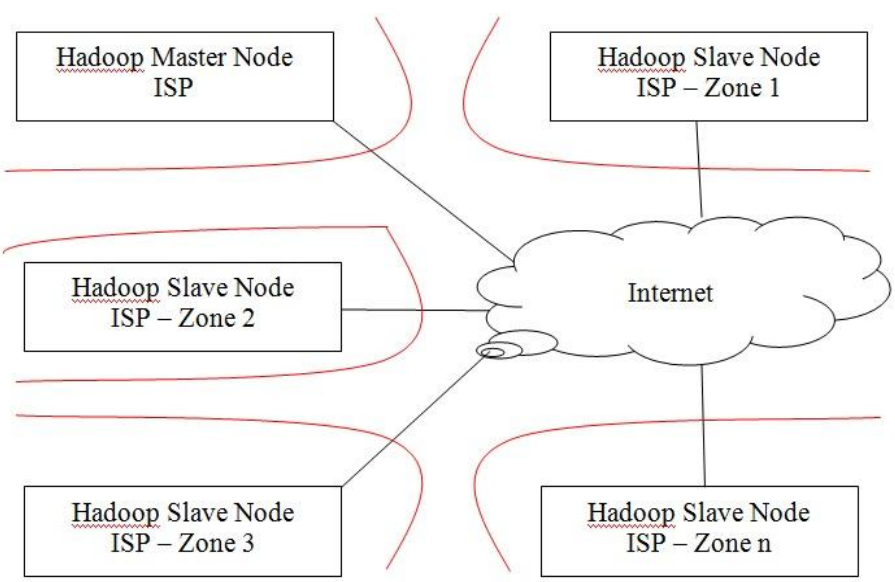

Figure 3 Hadoop Design with ISPs

3. ADVANCES IN NETWORK TRAFFIC ANALYSIS

This section discusses the recent era of network traffic analysis. Some of the most recent and widely used network traffic analysis methods are discussed in this section. Following types of analysis can be done on network traffic.

3.1 Traffic Management

At various levels, Internet traffic can be analyzed for quality of service, experience, bandwidth, throughput, security, privacy related analysis. Such analysis can be classified into four categories: Traffic Measurement / Monitoring, Traffic Characterization, Traffic Classification and Traffic Identification. The analysis can be either in online mode or in offline mode. Online analysis is mainly based on analysis of live traffic to detect security threats by analysis of packets and traffic pattern where as offline analysis is mainly based on analysis of generated logs. Offline analysis refers to the furnishment of various reports like usage reports for billing etc. in recent years, the network traffic analysis has become very required feature. From individual packet analysis to analysis of Internet usage of every organization,every analysis procedure has its own importance. Some extremely application specific analysis such as detection of DDOS attack to some addon analysis features such as users at risk, client reputation score are introduced. Such analysis help in spreading awareness to the users as well as to identify malicious users and those who are at risk [8,9].

3.2 Reputation Tracking

Fortinet's Foritgate firewall comes with a very interesting and useful feature called client reputation tracking. Figure 4 shows how Fortigate keeps track of information related with five categories. The administrator has rights to decide importance of each of this categories with a scroll bar as shown in Figure 4. The four important categories are: Application Protection, Intrusion Protection, Malware Protection, Packet Based Inspection and Web Activity [8].

Figure 4 Fortigate's client reputation tracing[8]

3.3 Application Sensors

The FortiGate recognizes traffic generated by various applications. Application control sensors are used to specify what actions you want to take if traffic is generated by specific applications. These actions could be in the form of either block or monitor or allow. Figure 5 shows an example of application sensor[8].

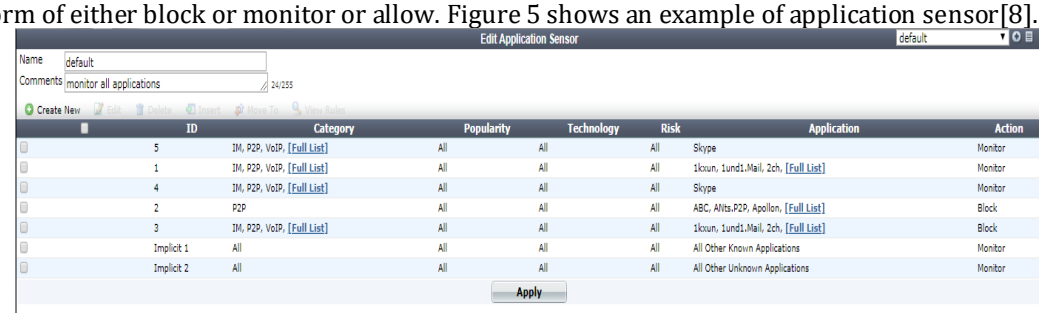

Figure 5 Fortigate's Application Sensors[8]

3.4 Data Leak Prevention

Fortigate's DLP Data Leak Prevention allows us to analyze what data is going out from our network. We can decide some rules in the form of patterns to block or log or allow data. DLP sensors can be created for this reason. These rules can be created with the help of some well known characteristics such as file name, file size, file type. For advance rule specification and for generalization of rules, it is also possible to specify rules in the form of regular expressions. These rules can be set in two ways for Messages as shown in Figure 6 and for Files as shown in Figure 7 [8].

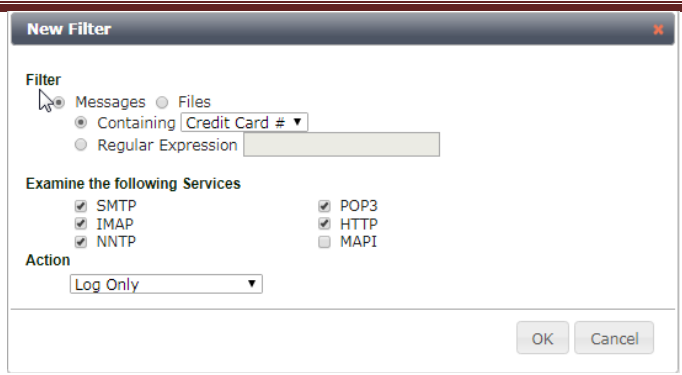

Figure 6 DLP for Messages[8]

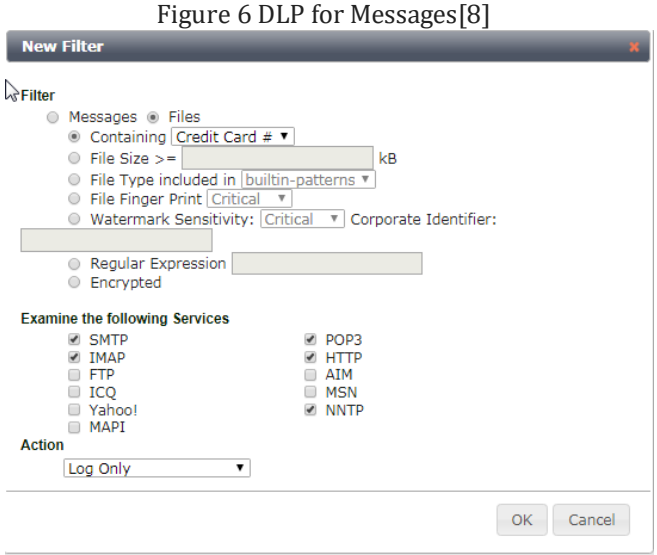

Figure 7 DLP for Files[8]

3.5 Other Features

Fortigate's other important features include web filters to define various policies which will allow or monitor or block websites as per urls. A set of categories for urls are shown in Figure 8. Along with the antivirus features, to detect intrusions, Intrusion protection has a set of IPS signatures as shown in Figure 9[8].

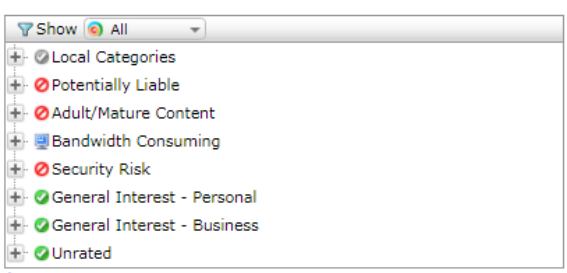

Figure 8 Web Filter Profiles[8]

Application Name	Severity	Target	
3Com.3CDaemon.FTP.Server.Buffer.Overflow		Server	Windows
3Com.3CDaemon.FTP.Server.Information.Disclosure		Client	Windows
3Com.Intelligent.Management.Center.Information.Disclosure		Server	Windows
3Com.OfficeConnect.ADSL.Wireless.Firewall.Router.DoS		Server	Linux
3ivx.MPEG4.File.Processing.Buffer.Overflow		Client	Windows
3S-Smart.CODESYS.Gateway.Server.Directory.Traversal		Server, Client	Windows
3S-Smart.CODESYS.Gateway.Server.DoS		Server, Client	Windows
3S-Smart.CODESYS.Gateway.Server.Heap.Buffer.Overflow		Server	Windows
3S-Smart.CODESYS.Gateway.Server.Integer.Overflow		Server	Windows
3S-Smart.CODESYS.Gateway.Server.Memory.Access.Error		Server, Client	Windows
3S-Smart.CODESYS.Gateway.Server.Opcode.Heap.Buffer.Overflow		Server, Client	Windows
3S-Smart.CODESYS.Gateway.Server.Stack.Buffer.Overflow		Server, Client	Windows

Figure 9 IPS Signature Sample[8]

4. HADOOP BASED CLIENT REPUTATION SCORE CALCULATION

4.1 Purpose

Fortigate's other important features include web filters to define various policies which will allow or monitor. We use firewall logs which are generated for users. These logs are having data of URLs visited / requested by a user along with its classified category. We analyze this data to calculate a client reputation score [10,11,12]. A client will be scored in the range of 0 to 100 where 0 is the least and 100 is the best score. We use three levels of categorization which are listed below.

1. HTTP Vs HTTPS URLs
2. URL classified labels
   - General Interest
   - Adult / Mature Content
   - Potentially Liable
   - Security Risk
   - Unrated
3. Firewall Action (Allow Vs Block)

4.2 Parameters

Table 1 shows the parameters which are calculated by our hadoop map reduce program from the firewall logs. Table 2 shows the weights assigned to each of these parameters.

Sr.	Parameter	Purpose
1	UserID	User identification
2	No_URLs	No. of URLs are requested. We consider requested URLs to analyze those URLs which are blocked by firewall.
3	No_HTTP	No. of HTTP URLs visited. This will include a few non HTTP based resources too. It is considered inside this category due to possibility of very small number of such requests.
4	No_HTTPS	No. of HTTPS URLs visited.
5	No_Allow	No. of URLs allowed by firewall.
6	No_Block	No. of URLs blocked by firewall.
7	No_General	No. of URLs classified as general interets.
8	No_Adult	No. of URLs classified as adult content.
9	No_Liable	No. of URLs classified as potentially liable.
10	No_Unsecure	No. of URLs classified as security risky.
11	No_Unrate	No. of URLs which are not classified by firewall.

Table – 1 Parameters

Sr.	Parameter P_i	Weight W_i
1	No_HTTP	2
2	No_HTTPS	3
3	No_Allow	4
4	No_Block	-4
5	No_General	3
6	No_Adult	-3
7	No_Liable	-2
8	No_Unsecure	-1
9	No_Unrate	1

Table – 2 Weights for parameters

4.3 Reputation Score Calculation

The client reputation score is the weighted sum of all values of parameters and corresponding weights. Here the parameter values are taken with reference of the amount of their involvement with reference of the total number of URLs visited. The formula to calculate client reputation score for use n is as below,

$$f(n) = \sum_{i=1}^{i=9} \left(\frac{\text{Pi}}{\text{No\_URLs}}\right) * \text{Wi}$$

$$Score(n) = f(n) * 10$$

4.4 Analysis

We have implemented above system with the help of map reduce facility of single node cluster design with cloudera hadoop environment. We have also evaluated the performance for a large number of users with different profession, age and gender. Our algorithm works best in calculation and bounded by [0,100]. 0 as the lowest bound and 100 is the highest bound. Some of the cases are listed below.

Sr.	User1	User2	User3	User4	User5	User6	User7	User8
No_URLs	10000	10000	30000	30000	50000	50000	25000	25000
No_HTTP	8000	5000	23000	14300	41800	34000	0	25000
No_HTTPS	2000	5000	7000	15700	8200	16000	25000	0
No_Allow	8500	7000	24000	27400	47100	41000	25000	0
No_Block	1500	3000	6000	2600	2900	9000	0	25000
No_General	8000	4500	19500	25050	37200	38700	25000	0
No_Adult	1000	2400	2000	2300	7600	4000	0	25000
No_Liable	0	500	2500	300	2200	3000	0	0
No_Unsecure	0	400	2000	2300	1200	2600	0	0
No_Unrate	1000	600	4000	50	1800	1700	0	0
Score	72.0	46.5	62.8	80.0	74.0	68.2	100	-50=0

Table – 3 Analysis

Table 3 shows some of the results of our work. Due to limited length of the paper, it is not possible to include each and every scenario in this paper but we have tried to show 8 completely different scenarios in terms of number of visited URLs and type of URLs. Out of these 8, to test accuracy of our work, we have generated two dummy scenarios to depict a user with best score of 100 (user 7) and a user with least score of 0(user8). As per our discussion the score is in the range of 0 to 100 so for user 8 it is rounded to 0 instead of considering negative value.

CONCLUSION

This research work is carried out to address the problem of analysis of network traffic log for Internet users. Due to computational limitations in terms of limited processing capabilities, memory and storage capabilities it is difficult to perform complex analysis of user logs in a traditional centralized manner. Further to it, the usage of the Internet is growing rapidly leading to the large number of web resources are being added to the Internet at the same time a large number of users start accessing the Internet every day. To cop up with the need of processing such extra huge amount of data, we have proposed a Hadoop based system to calculate a client reputation score. Just like a student's performance can be predicted based on the single value called his degree percentage, our goal is to analyse users' internet access pattern with a single value called its reputation score. We have implemented this work with the map reduce environment of Hadoop aligned to the firewall logs format. We have found our system is much useful for any organization to keep eyes on the usage of their users. Based on the reputations, an organization can immediately take actions to block users accessing the Internet without digging into what resources they are using.

FUTURE WORK

The work presented in this research is based on analysis of high volume network data of a user to calculate his/her reputation score. The purpose is to let organizations know how genuine and safe their users and their Internet usage are. We have included various categories which are maintained by most of the current firewalls available in the market. Further to the research, we can calculate client score based on volume of data they use. In this work we have purposely not included that factor because it is more important what data is accessed rather than what size it has. But we can also enhance this system by analysis which protocols are used at lower layers in the network. We can also do analysis on time such as what resources are accessed when. The algorithm is tested on single node Hadoop cluster. It can be implemented in real life network in a distributed multi node Hadoop cluster too.

REFERENCES

1. Managing Internet and Intranet Technologies in Organizations: Challenges and Opportunities, Subhasish Dasgupta - July 2000
2. www.statista.com - Digital population in India as of January 2018 (in millions).
3. www.statista.com - Number of internet users in India from 2015 to 2022 (in millions).
4. Cisco Visual Networking Index: Forecast and Methodology, 2016–2021

5. Hadoop, http://hadoop.apache.org/.
6. Shvachko, Konstantin, et al. "The hadoop distributed file system." Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. IEEE, 2010.
7. J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Cluster, OSDI, 2004.
8. www.fortinet.com
9. Lee, Yeonhee, and Youngseok Lee. "Toward scalable internet traffic measurement and analysis with hadoop." ACM SIGCOMM Computer Communication Review 43.1 (2013): 5-13.
10. Lee, Youngseok, Wonchul Kang, and Hyeongu Son. "An internet traffic analysis method with mapreduce." Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP. IEEE, 2010.
11. Big Data Analytics with R and Hadoop, Vignesh Prajapati-Packt Publishing Ltd – 2013
12. Cloudera: Third Edition, Gerard Blokdyk, Create Space Independent Publishing Platform, 2017