Network Surveillance: Detecting Anomalous Activities in the Network using Big Data Technologies

Pratiksha Ghate[1], Prof. Mirza M. Baig[2]

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
J D COLLEGE OF ENGINEERING AND MANAGEMENT
NAGPUR, MAHARASHTRA

ABSTRACT

Abstract— Telecommunication companies receive massive amounts of data providing lots of information about the customers and network. This data, both structured and unstructured, when analyzed can reveal deeper insights into customer behavior, their service usage patterns, preferences, and interests in real-time. The ongoing digitalization in every field is making enterprises vulnerable to cyber attacks. The traditional application and storage architectures are not designed for storing and processing such 'Gigantic data' due to which a lot of useful information is lost. Analytics can be thought of as a weapon in maximizing cyber resilience. That is where big data analytics comes into play.

This project addresses the issue of handling the large volumes enterprise data using Big Data - Hadoop and applying analytics to detect patterns in anomalous user activities based on their internet browsing and usage.

Keyword: - *RDBMs, Big Data, Hadoop Distributed File System, HDFS, HIVE, Tableau*

1. INTRODUCTION

With the rapid adoption and continually increasing growth of smart phones and numerous other inter-connected network devices, customer's way of interacting and consuming telecommunication services has completely changed. With latest devices, such as Apple's iPhone and Google's Pixel, the web-surfing experience has gotten even better than what it used to be on traditional desktops and laptops. This has led to increased consumption of video intensive and interactive applications like YouTube and Netflix, social networking and sharing platforms like Facebook, Snapchat and infinite others. With parallel advancement in telecommunication sector, enterprises have been highly successful in delivering the services that their customers have been demanding. With customer's sole expectation to have latency-free data speeds to access digital media, these companies have been more than just successful in delivering those. This increase in network performance has triggered a dramatic growth in data traffic. This high volume, high velocity, once considered as useless, is now tagged as the source of truth for further improvement in these network services. These network related data sets are largely unstructured when it comes to telecom companies and can be broadly categorized into:

- Customer Loyalty Data – can be utilized to know the customer churn, retention and acquisition
- Network Data– can be utilized to further improve the existing services
- Security Data – can be utilized for fraud detection to promote cyber-security

This information can be stolen or used in a way to disrupt or shut down any existing services. Detection and prevention of such threats is of utmost importance to an individual or an organization. Effective measures can be taken to curb the menace caused by such threats.

This data when analyzed could be helpful in extracting insights that could trigger the improvement in network performance to meet the ever-increasing demand for better services. This is possible only when the data could be first collected in a way that is repeatable, scalable and actionable. With exponential growth forecasted for smart devices leading to a heavy increase in data traffic, it has been a challenge for ISPs to collect and analyze this data. The traditional infrastructure that most of the enterprises today have inefficient data storage, ineffective processing frameworks and even lack rich sets of analytical capabilities. Legacy system frameworks need to be replaced by Big data frameworks in order to completely and sufficiently utilize the power of big data.

With such a scaling data, threats and attacks are growing in line with the increase in complexity. Enterprises need to reiterate on their concepts on cyber security. They have to move towards the 'PDR paradigm - Prevent Detect and Respond'. Considering the fact that millions of records flow into the Telecom databases every second, there is a need to perform accurate analysis. Big data technologies provide the unique ability to segregate data and computation which is a big improvement over the traditional tools.

The method proposed uses the Big Data architecture which is based on open source Hadoop[1] and uses some of the eco-system components to cater to the requirement. This approach uses the technique of analyzing the network logs for anomalous activities using Big data technologies. The analysis of network security is carried upon the data captured in the Squid Proxy log files[2, 3]. Squid is one of the widely used HTTP proxy implementations which is extremely flexible and customizable. The sample dataset is collected from an open dataset provided by SecRepo: Samples of Security related Data[4]. The extracted data is exported to txt, preprocessed and uploaded onto HDFS environment. This data is then imported to HIVE[5, 6, 7] for further processing. Analytics is performed using HIVE queries. Finally to make complex ideas look simple, data visualization is done using Tableau[8].

The rest of the paper is organized as follows. In section 2, the related work in the area of network analysis is discussed. Section 3.1 presents the proposed model. Section 3.2 briefs about the results and experimental analysis and the final section 4 concludes the paper.

2. LITERATURE SURVEY

There are several research papers that compare the performance evaluation of storing and processing huge amount of data between the big data technologies and the traditional enterprise systems.

Nattawat Khamphakdee et al. [9] proposed a system that analyzes and processes the big network traffic data. The proposed Hadoop-based traffic querying and analyzing system handles the TCP, ICMP, and UDP analysis of the big network traffic data. They have also compared the response query times of MySQL with Hive for certain queries.

A. Fuad et al. [10] compared the data model from GroupLens Research Project by executing simple queries to show how Hive or Pig is faster than MySQL cluster. The result shows that, Hive is able overcome both Pig and MySQL cluster on the low-cost hardware environment.

Vibha Bhardwaj, Rahul Johri. [11] in their paper have explored the issues and challenges of big data tools currently used to implement and analyze big data. They have mentioned many applications where Big Data can be or is already deployed as a solution. A broad comparison of various data mining techniques have also been stated.

Sayalee Narkhede et. al. [12]This system applies Hadoop MapReduce programming model for analyzing web log files so as to get hit count of specific web applications. The stored log files are evaluated using Map and Reduce function. Experimental results show the hit count for each field in log file.

The proposed system is based on investigating the web logs and to find the patterns and behavioral trends in line with security. The logs are analyzed to summarize the usage patterns of various users. Different visualizations will provide an easier understanding of the use cases.

3. NETWORK SURVEILLANCE USING BIG DATA TECHNOLOGIES

3.1. Proposed Model

The Proposed Model involves the 5 phases as shown in Fig 1.

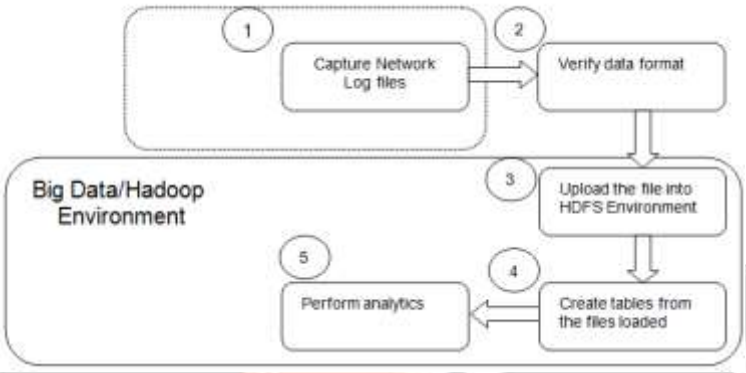


Fig -1: Proposed Model

First phase involves capturing the network log files. The application log files are captured from the Proxy servers of Local ISPs. These files serve as input to the system. Table 1. describes the native format for Squid Proxy files. In the second phase, verification of data format is done. After the files are received, they are cleaned and preprocessed before uploading it to the Hadoop Environment. The log files received are in simple text like format. They are first preprocessed to add delimiter '|' so as to separate out the fields contained in the log files. In the third phase the file is then loaded into HDFS environment for further processing and analytics. Following tables are created in the fourth phase into which the files will be imported.

1. the log file in external Hive table named - *accesslog*
2. suspected words file into *suspectedWords* table

These tables are created using Command Line Interface. Table 1. shown below describes the schema of the table created for importing the log files. In the final phase analysis is done using Hive queries to identify various browsing activities and patterns. To create multi-faceted views, Tableau is used for the purpose of Data Visualization.

The imported log files are compared against suspectedWords table. The schema for suspected words is as shown in Table 2. The suspectedWords table contains a list of the suspected words and their categories. The suspected words can fall into various categories like Terror, Piracy, Adult, Drugs, Others etc. Evaluating various types of queries help gain an insight into trending categories and user interests which thus help in taking necessary actions.

Table -1: Accesslog table schema with field description.

Field	Description	Datatype
time	A Unix timestamp as UTC seconds with a millisecond resolution.	bigint
elapsedtime	How many milliseconds the transaction busied the cache	int
remotehost	The client IP address	string
code_status	The cache result of the request contains information on i. the kind of request, ii. how it was satisfied, or in what way it failed.	struct<code:string,status:int>
bytes	The size is the amount of data delivered to the client.	int
method	The request method to obtain an object.	string
url	This column contains the URL requested	string
rfc931	May contain the user identity for the requesting client.	string
peerstatus_host	Consists of three items: i. Any hierarchy tag may be prefixed with TIMEOUT_ ii. A code that explains how the request was handled. iii. The IP address or hostname where the request (if a miss) was forwarded.	struct<peerStatus:string,peerHost:string>
type	The content type of the object	string

Table -2: Suspectedwords schema with field description

Field	Description	Datatype
srNo	Serial no.	int
word	Suspected word	string
category	Category of suspected words	string

3.2. Experimental Results and Analysis based on Hive Queries and Tableau visualizations.

Web logs can be analyzed to see what is currently trending, the usage patterns and the user interests. Interesting analytics were observed after querying the database.

Following use cases were created to form base of the proposed system

- Finding out the Top 10 active users

  This query tells us how many websites and webpages a user has visited. A user can visit different webpages on the same site or different websites altogether. The query can be further drilled down to see the no. of clean or suspected sites a user has visited.

  The measure 'Suspect Check' signifies two aspects:

  - Clean - URLs without suspected words
  - Suspect - URLs containing suspected words.

Table 3. Top Active Users arranged descending by webpages visited

Remote Host	Suspect Check	Distinct count of URL	Distinct count of Resolved URL
10.105.51.137	Clean	1,541	133
	Suspect	3	2
10.105.37.65	Clean	1,174	138
10.105.37.166	Clean	1,089	264
	Suspect	26	4
10.105.35.227	Clean	951	102
	Suspect	2	1
10.105.23.145	Clean	916	93
	Suspect	25	4

Table 4. Top Active Users arranged descending by webpages visited

Remote Host	Suspect Check	Distinct count of Resolved URL	Distinct count of URL
10.105.37.166	Clean	264	1,089
	Suspect	4	26
10.105.37.65	Clean	138	1,174
10.105.51.137	Clean	133	1,541
	Suspect	2	3
10.105.37.171	Clean	134	476
10.105.33.225	Clean	125	382
10.105.35.227	Clean	102	951
	Suspect	1	2

From the above two tables it is interesting to see that users who visited maximum WebPages is not the same as who visited maximum websites.

- Find out the Top 10 sites browsed by the users.

  The query displays the results for the no. of the times the website was visited along with the no. of unique visitors for that website. This gives us an idea about the size of the audience for the websites.

Table 5. Top websites arranged descending by unique visitors

Resolved URL	Distinct count of Remote Host	Count of Remote Host
download.windowsupdate.com	169	1,861
Null	166	3,832
download.microsoft.com	152	1,626
us.i1.yimg.com	99	11,310
windowsupdate.microsoft.com	98	1,311
us.bc.yahoo.com	97	1,022
us.js2.yimg.com	89	1,206
view.atdmt.com	87	464
mail.yahoo.com	85	183
us.mcafee.com	84	369
us.a2.yimg.com	73	1,222
update.microsoft.com	71	320
spe.atdmt.com	66	160
www.google.com	65	1,033

- User Activity on Suspected Websites

  Elapsed time and the total no. of bytes downloaded from a website can be analyzed to see the amount of download happening where high levels of download might mean higher user activity.

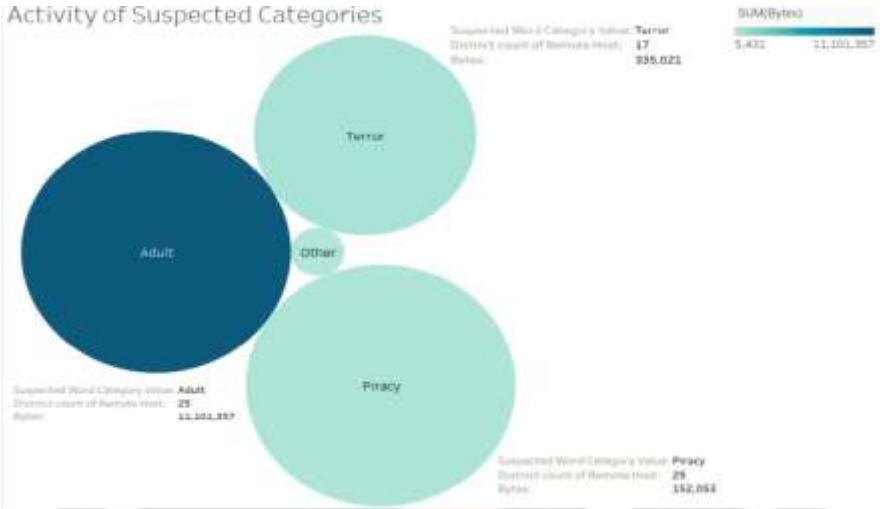
Table 6. User Activity on Suspected Websites

Remote Host	Resolved URL	Bytes	Elapsed Time
10.105.23.174	[illegible]	9,140,714	776,363
	[illegible]	593	1,749
10.105.37.12	photos.adultfriendfinder.com	140,640	36,744
	www.leslieshore.com	88,912	35,731
	adultfriendfinder.com	79,550	10,203
	www.videhoe.com	37,590	4,196
	graphics.adultfriendfinder.com	33,825	1,729
	us.f3.yahoofs.com	19,743	3,443
	[illegible]	13,496	2,132
	banners.adultfriendfinder.com	7,627	2,282
	Crocko.com	1,489	3
10.105.37.166	photos.adultfriendfinder.com	198,676	35,233
	graphics.adultfriendfinder.com	64,235	7,435
	www.ua-torrent.net	2,689	4,158
	ec1.images-amazon.com	1,703	6
10.105.23.145	photos.adultfriendfinder.com	68,707	18,528
	graphics.adultfriendfinder.com	67,849	3,274
	us.f3.yahoofs.com	38,033	5,836
	adultfriendfinder.com	662	15,703

- Activity of Suspected Categories

  This use case demonstrates the kind of URLs the users are accessing more. The URLs can be grouped into various categories like Terror, Piracy, Adult Content etc.. The data can be analyzed by finding out how much content is being downloaded from such suspected categories. It can be used to see how many different users are trying to access such websites falling under suspected categories.

Chart 1. Activity in Suspected Categories

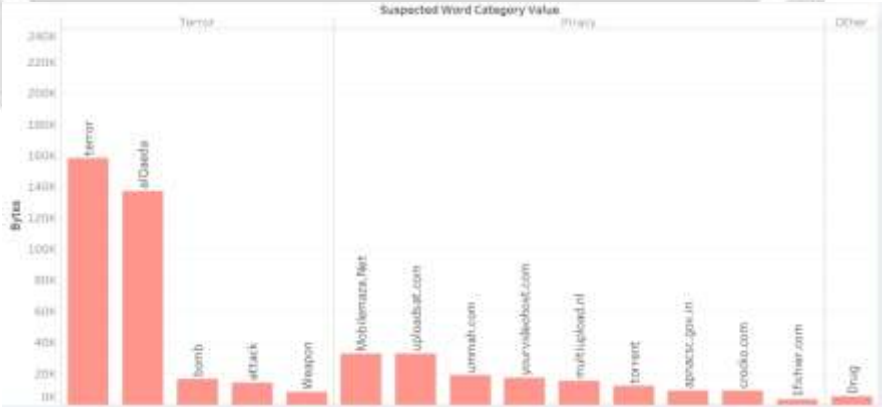


The above visualization shows that the Adult category URLs were the most accessed in the downloaded bytes category followed by Piracy, Terror and so on, whereas Adult and Piracy bubbles had same size of audience followed by Terror. This gives an insight into the size of audience for a particular category.

- Activity of Suspected Category words

  The suspected category Use-case can further be drilled down to see what specifically people are accessing under that particular category.

Chart 2. Activity in Suspected-words categories



Have excluded the Adult category from the results for the purpose of visualization in the graph, which exceeded in the content download followed by Terror, so as to show the impact of other words in different categories.

- Blocked Access (per User)

The result returns the no. of times a user tried to view a blocked URL for which he might not have the necessary permissions, or doesn't have credentials for the same. Analysis can be done to see if the count of no. of blocked sites exceeds some threshold value which might mean attention is required for such users and sites.

4. CONCLUSIONS

The amount of data that telecom companies are receiving is very large. Effectively analyzing such exponentially growing data with traditional storage and analytical solutions is a big challenge. Enterprises cannot effectively use the information available to them to gain deeper insights. Thus the proposed system handles the scalable data with ease by using the Open Source Big Data framework. Analysis done using this technology not only helps in digging the data better but also provides a good insight into the usage patterns and user interests. Visualization totally accelerates the time taken to understand the patterns generated using the Hive queries. This system can surely benefit an organization in monitoring and limiting the anomalous activities going on in the network.

5. ACKNOWLEDGEMENT

Many thanks and appreciations to the institution for guidance and supervision in completing this project. We thank all the people who directly/indirectly provided us with their constant support and co-operation that assisted in the Research.

6. REFERENCES

[1] Jonathan R. Owens, Jon Lentz, Brian Femiano. Hadoop Real World Solutions Cookbook. 2013 Packt Publishing

[2] Squid Cache Wiki's webpage : http://wiki.squid-cache.org/Features/LogFormat#squid_result_codes

[3] Squid The definitive guide avaliable at
http://etutorials.org/Server+Administration/Squid.+The+definitive+guide/Chapter+13.+Log+Files/

[4] SecRepo: Samples of Security related Data available at :http://www.secrepo.com/squid/access.log.gz

[5] Dayong Du. Apache Hive Essentials , 2015 Packt Publishing

[6] Dean Wampler, Edward Capriolo, and Jason Rutherglen. Programming Hive. 2012 O'Reilly Media

[7] Language manual for HIVE + UDF available at:
https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF

[8] Tableau tutorials available at, https://www.tableau.com/learn/training

[9] Nattawat Khamphakdee, Nunnapus Benjamas and Saiyan Saiyod "*Performance Evaluation of Big Data Technology on Designing Big Network Traffic Data Analysis System*" Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems, 2016 Joint 8th International Conference on 25-28 Aug. 2016

[10] Ammar Fuad, Alva Erwin and Heru Purnomo Ipung "*Processing performance on Apache Pig, Apache Hive and MySQL cluster*" Information, Communication Technology and System (ICTS), 2014 International Conference on 24 Sept 2014

[11] Vibha Bhardwaj, Rahul Johari.'Big Data Analysis:Issues and Challenges'. 2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)

[12] Sayalee Narkhede, Trupti Baraskar, Debajyoti Mukhopadhyay. Analyzing Web Application Log Files to Find Hit Count Through the Utilization of Hadoop MapReduce in Cloud Computing Environment: IT in Business, Industry and Government (CSIBIG), 2014 Conference