Security Visualization on Big Data

Waqar Ahmed
MS Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology
90 & 100 Clifton
Karachi -75600
cpl_waqar@hotmail.com

Mr. Uzair Hashmi

Head of I.T Security Karachi Stock Exchange
uzair.hashmi@kse.com.pk

Abstract—— **IN 2012 2.5 QB OF DATA (18 ZEROS AFTER 1) WAS GENERATED WORLDWIDE KNOW SO FAR. EVERY DAY DATA CREATION SIZE IS BECOMING BIG-TO-BIGGER THAN WAS SEEN BY EVERYONE SINCE THE BEGINNING OF HUMANKIND. BRIEF DATA GENERATION HISTORY IN CATEGORICAL FASHION IS AS FOLLOWS:**

- **2.5 Petabytes: Data flowing through Walmart's transactional databases.**
- **Consumers spend $272,000 on Web shopping /day.**
- **Apple receives around 47,000 app downloads /minute.**
- **On Facebook, Brands receive more than 34,000 "likes" /minute.**
- **144.8 billion Email messages per day**
- **On Twitter 340 million tweets per day**
- **On Facebook 684,000 bits of content per day**
- **3,125 new photos uploaded on Flicker per minute.**

As data size increased so are security threats, which comprise unauthorized modification/ alteration of big data. Conducting security visualization manually on such a large-scale data is beyond compression. Therefore, we need some automated easy to use, time saving technique that can give comprehensive results, which can help to track the integrity of big data.

Keywords— Big Data, Security Visualization

I. INTRODUCTION

In today's cyber world, data created at enormous rate, and so are security threats. Hackers restlessly attack with various methodologies on potential organizations. Moreover, it becomes more and more difficult for security personnel to analyze security threats and pattern of traffic on such huge size of data specially when its motion by applying manual strategy. When it is matter of data security, security analysts need some effective visualization techniques and tools that can increase productivity of security analysts tremendously. Conducting security visualization manually on such a large-

scale data is beyond human comprehension. Therefore, we need some automated easy to use, time saving technique and tools that can give comprehensive results, which can help to track the integrity of big data.

We collect more and more data from our stock of applications. This large size, multi-type and multi-format of data gets very difficult if not impossible for comprehension if visualized manually. Therefore, we need some visualization technology that can help security professionals to manipulate, analyze and finally devise remedial action to any threat found in our information assets, and protect confidentiality, integrity and availability of data.

In this research report, I will materialize the various data analysis tools, with comparison eye to come up with most effective security visualization tools. I will also highlight the limitations of these tools. Finally, I will come up with the new tools that are effect, robust, economic in resource utilization and time saving so that productivity of security analysts will increase tremendously.

Using big data visualization we can acquire great spatial understand that what is happening currently on our network. We can make our visibility sharper on patterns of data flowing through organization's infrastructure. Those data pattern once define can help security analysts at great extent to identify timely and effectively the emerging threats and attacks patterns fired by hackers. Thus, security professionals will be fully aware of their network/ system's health, which will help them to devise countermeasure to those attacks and eliminate the threats to possible minimum level. On other hand if this process of large size data visualization and analysis is conducted manually or with conventional methods and tools its will get their nerves and it is likely possible that they can miss the target.

II. WHAT IS BIG DATA?

2.5 QB (quintillion bytes) data generated every day worldwide. Organizations generating this huge size of data includes; climate information gathering sensors, social media stuff, videos hosting websites, purchase and sales transactions, financial institutions' data, cellular phones' data, and GPS signal data are the few names. This enormous sized data is called as big data.

There are four dimensions of big data as follows:

1) *Volume:* Enterprisers are facing ever-growing data flood of various types that nicely and easily occupies large amount of storage space like in terabytes or petabytes. Almost 13 TB of Tweets are created every day for analysis and

improvement of product NADRA's manipulation of 18 billion people's multidimensional data for various informatics

2) Velocity: Sometimes two minutes are too late and here comes preciousness of time! Time sensitive processes like catching fraudulent transactions, big data must be used as its streams into your organization with objective of maximizing the value organization.

Inspect five million trade events created each day for recognition of potential fraudulence Examining 500 million call detail records every day in real-time to anticipate customer churn faster

3) Variety: Big data is mixture of all and every kind of data i.e. structured and un unstructured data such as video, audio, text, still pictures, sensor data, clicking streams log containing files and many more types.

III. CURRENT TECHNOLOGY LIMITATION OF SECURITY ANALYSIS

Organization around the world uses various IDSs to monitor their networks' security but these sorts of tools have tragic flaws. When logs exceed in Giga bytes then it is likely that human may make error. There are several dozen tools both proprietary and open source that are used for security. However, in following I have shown the results the basis of experimental work that will reveal the limitations of these security tools. World famous tools used for experiments are as follows:

A. *Unification of NOC And SOC:* In last decade, security professionals tried to put together the Network Operation Center and Security Operation Center to cope with the ramping security analysis problem, but eventually it failed. Since most of NOC and SOC tools had interface issues, therefore this model also could not run successfully.

B. *Application Monitoring:* Then organizations started a new method; collecting logs data for their infrastructure purposes and also web analytics directly from application, they started stored at one place and started pumping-in to Security Information Managements (SIMs) but the result was chaos! It just blew up because it was too much data and schema did not worked and whole lot of problems came in. Therefore, it practically went void.

C. *IDS:* Snort is intrusion detection system (IDS). It is capable of performing real-time traffic analysis and if deployed with database system it can log the traffic of host and network as well. It freely available under the GNU general pubic license (GPL). By making slightly variation in deployment, it is also used as network intrusion detection system (NIDS). Snort is open source, and provides opportunity to hook with third party modules like

D. Barnyard2 etc. It working methodology is based on the signatures pattern matching. The security professionals define these signatures.

Fig1. Snort data capture

Limitations:

To understand the limitations of Snort we have to understand first its working mechanism. Snort receives packets from some packet-capturing driver i.e. WinPcap etc., then that packet is matched with predefined signatures, which are laid by security professionals. Now if somehow received packet and predefined signature mismatches due to encrypted shape of malicious payload then Snort is going to ignore that packet. In addition, that packet may part of some hacker's hacking attempt data.

Security analyst totally rely on the alarm generated by the snort, in aforementioned scenario data packets containing some malicious payload may go "false positives"

As show in figure 2.1 is the Snorts data capture, in this format it is very rare for security analyst to visualize and comprehend the security especially when this log is in size of gigabytes.

Third party visualization tools known as Barnyard2 may be integrated with Snort but it only shows the volume of traffic: not complete picture the security scenario. As show in figure 2.2, Snort is partially unreliable for security.

E. *Wireshark:* Wire shark is basically protocol analyzer but it has been also used by security professionals to monitor what is going inside their network infrastructure.

The software is free and available, can run on almost on all operating system platforms like all flavors of Linux, Windows, Mac OS, Unix, and Solaris and several other platforms.

Limitations

- Wireshark is packet-focused; it cannot generate data picture even in raw format.
- Creates piles of data, very difficult to handle.
- Wireshark's performance decreases tremendously when it captures large network files.
- The only way to trace the wanted (here malicious data) by color coding of the packets.
- • On small networks Generates almost 5-to-10 Mb data in one minutes (imaging how would it make logs in a day)
- By default, it will not warn security analysts about any malicious activity.
- It does not manipulate any outcome from captured data rather it will only measure the data.

- Wireshark do not capture 100% when data speed is high. Which increases the risk factor that some malicious data may go unnoticed
- Wireshark do not capture data between the frames, which is again black hole for security measures.

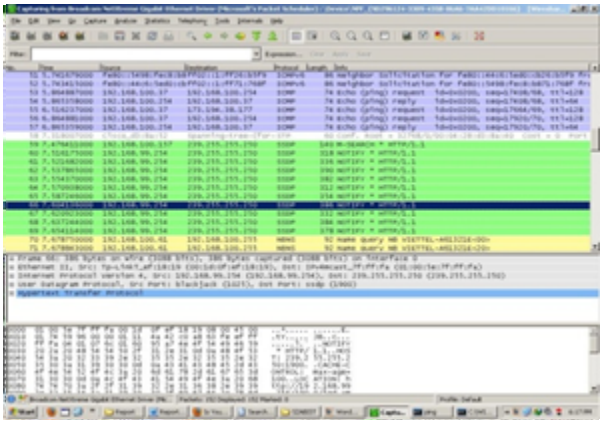

Fig 2. Wire shark data visualization

F. *Network Management SNMP And MRTG:* Multi Router Traffic Grapher (MRTG) is software that is used to only measure and monitor the network traffic in means of quantity. However, do not measure the quality (where security analysts need type of data etc.). There are many hundreds well know MRTGs, some open source and other paid. These sorts of tools lie under the category of network management system (NMS) whose focus is to check availability, performance and reliability.

MRTG works on the basis of Simple Network Management Protocol(SNMP).
Mostly MRTGs are proprietary; developed by/ for hardware vendors. Famous names includes

- SOLARWINDS
- ROUND ROBIN DATABASE
- Tobi Oetiker

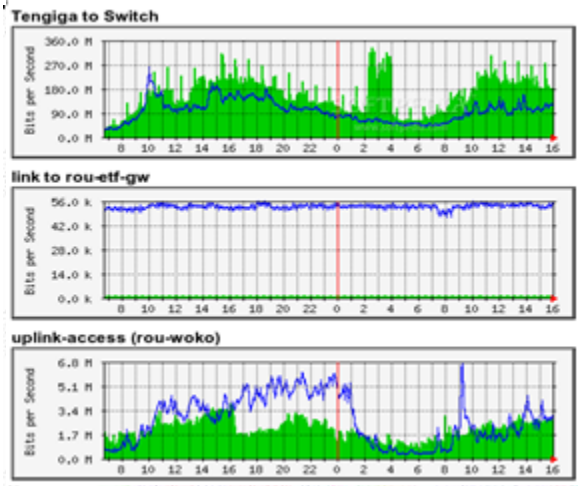

Fig 3. MRTG Output

IV. BIG DATA SECURITY ANALYSIS BY MODERN METHODOLOGY AND TOOLS

A. Mining Logs-Gaining Insight Through Visualization:

After trying, the techniques mentioned in chapter 2, which gave the result incomplete and unsatisfactory as those, cannot render the complete comprehensive results; here it is the new matured strategy with new tools that will be experience and explained in below.

1) Security Visualization Methodology: Figure 3.1 shows the devised methodology for security visualization of big data. Moving from left side to right side whole process of security visualization moves towards the maturity.

i. Starting from collection of log data files with the help of an effective packet capturing drive i.e. WinPcap, syslog, Pcap etc.
ii. Then if there are some irregularity regarding protocols, ports and format of the capturing that to be corrected by mean of tuning of the data-capturing driver or troubleshooting the hardware device (servers machines, routers, network infrastructure etc.).

Fig 4. Security Visualization Methodology

iii. Then those correction are to be saved(also for the sack of future work thus security professionals do not have to start from ground up) also share the settings/ troubleshooting to technical community so that if there is any flaw then that must be indicated.
iv. Because of those tuned setting security analyst are to create alerts, and keep observing those alters by mean of collecting more log data.
v. Correlate the alerts with each other to possible extent.
vi. Finding of patterns
vii. Interactive visualization by changing variables in log data.
viii. Then security analyst can dynamically issue queries.
ix. Anomaly detection
x. Thus, this whole process proceeds towards maturity level.

B. *Advance Tools (DAVIX):*

Linux is open source freely available security visualization suit of tools. The operating system is built on SLAX flavor of Linux. It is highly customized and modularized so that security analyst can tune it according to their needs and scenario. Also highly decorated help files are also included with each of several visualization tools so that professional can get help for those tools easily. Followings are among leading new tools contained in the operating system. DAVIX runs in live CD/DVD, USB, and VMware iso fashion. It is to be noted that 'Live' is because it is used for network forensics.

a) Log Management Architectures

Almost 90% of log data is based on famous among professionals is syslog tool. Also there are many proprietary tools i.e. opsec, netflow etc. Figure 3.2 shows the LMA process.

b) Normalization of Log Data

Now we need some parser, which can parse or search the queries of security analysts and manipulate and store the log data. In this case, consider the following figure 3.3 which normalization of log data is shown to make our queries effective
Normalization

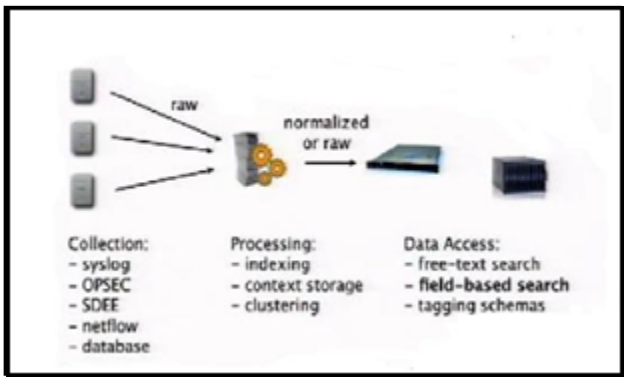

Fig 5 Log data

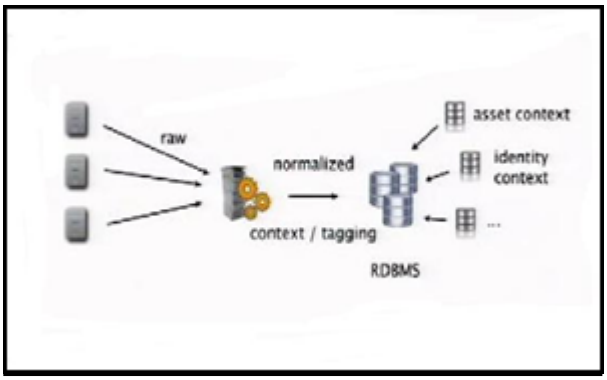

Fig 6. SIEM

c) Agents and Connectors

The log data is collect through agents and connectors. It is good approach because a piece of software is installed on end node machine, which helps in collecting data effectively. Figure 3.4 shows the details of agents and connectors

C. *SECURITY INFORMATION AND EVENT MANAGEMENT (SIEM):*

After log management stag in which we have mostly, raw data mostly connected to parsed data. This has very similar kind of structure as log management section has, but here we need all the data must be parsed and every data source must have connectors. If at this stage any data is not properly parsed of has no connector then generally SIEM is going to put it in error data.

a) Benefits of SIEM

- Parsed data enables real-time correlation and real-time statistics.
- SIEM also provides data augmentation(context) close to source
- Unified data access language over fixed set of fields
- It provides real time dashboards.
-

D. LOOGLY:

In figure 7, a complete picture of security visualization is revealed.

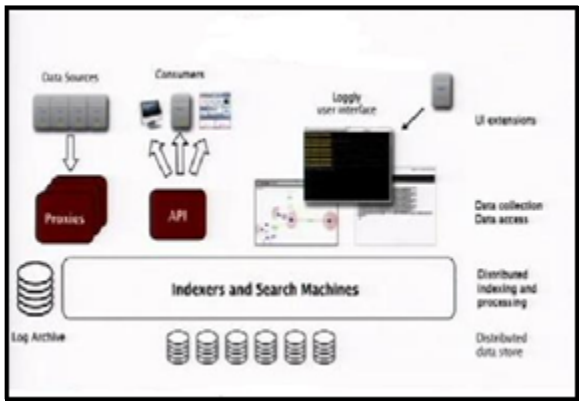

Fig 7 Loogly Model

I. COMPARISON TO TOOLS AND TECHNOLOGIES

Following chart provide clear picture of various discussed above tools and technologies and all pros and cons.

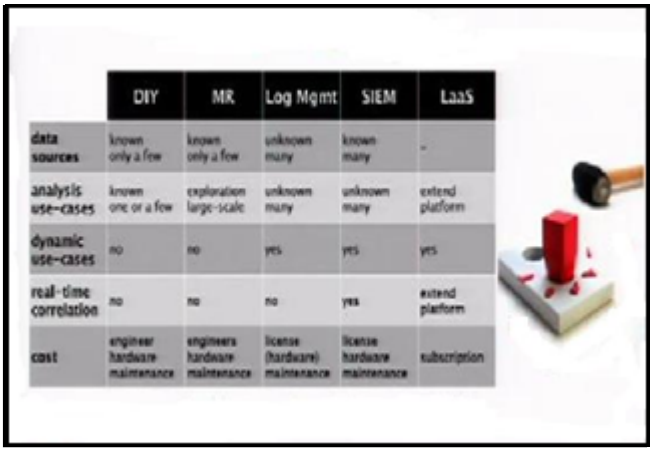

Fig 8 Comparison Chart

How effective Big Data easily can be analyzed by new tools!
Analyzing big data with the help of new tools will reveal effective and clear picture of traffic, which will turn big data into big action for potential organizations' business. It will also deliver the sharp results to the security analysts painlessly. New methodology and tools will display overall picture of the security which enables security analyst to filter results according to their needs.

E. AFTERGLOW

One of exciting security visualization tool available in DAVIX is afterglow, comparison between conventional tool output and of new tool is shown in figure below

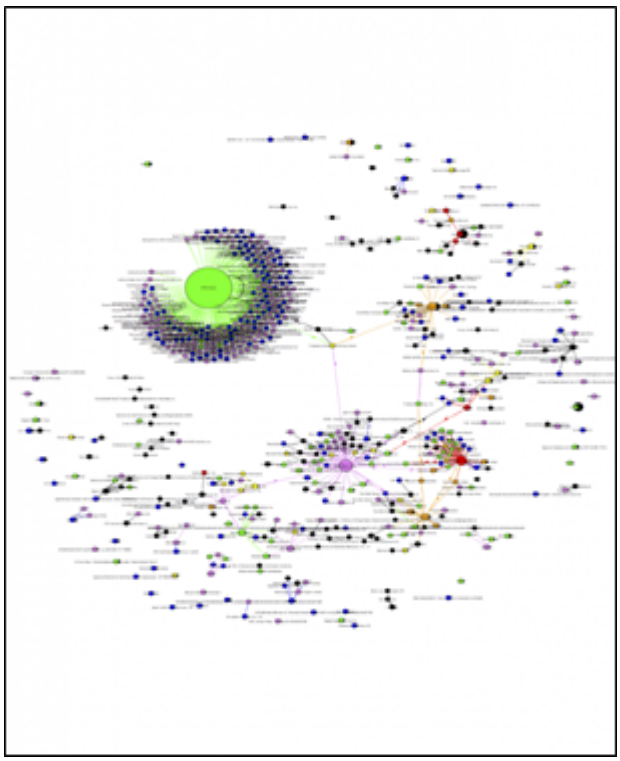

Fig 9 Afterglow output

V. CONCLUSION

In order to deploy an effective security visualization new tools following Dos and Don'ts are to be observed, violation of these finds may result in unsuccessful working of the tools
Logs must be collected carefully and centralized. Alerts must be defined on prioritized manner. Security analyst must frequently conduct forensics so that tools working remains updated. In deployment of DAVIX tools log formats are all over and if not documented properly then this may bring system ineffective. Log files must be standardized and documented.

References

[1]. Conti G. *Security Data Visualization*. No Starch Press, 2007
[2]. Marty R. *Applied Security Visualization*. Pearson Education, 2008.
[3]. Matějíček T. *SLAX 6*. http://www.slax.org
[4]. Monsch J. P., Marty R. *DAVIX Manual 1.0.1*. 2008. http://82.197.185.121/davix/release/davix-manual-1.0.1.pdf
[5]. Shneiderman B. *Keynote VizSec*. Boston: 2008
[6]. Shneiderman B. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization. *IEEE Visual Languages*. pp. 336 – 343. 1996.
[7]. Wireshark / tshark Manual http://www.wireshark.org/docs/wsug_html/
[8]. http://lcamtuf.coredump.cx/p0f.shtml Snort Manual
[9]. http://www.snort.org/docs/snort_htmanuals/htmanual_282/
[10]. AfterGlow Manual http://afterglow.sourceforge.net/manual.html
[11]. Network Forensic Analysis
[12]. Berkeley Packet Capture (BPF) and Related Technologies: AnIntroduction Alexandre Dulaunoy, November 29, 2012
[13]. DAVIX Manual Version 1.0.1 Authors: Jan P. Monsch, jan.monsch ät iplosion döt com, Raffael Marty, raffy ät secviz döt org database