Crimson Publishers
Wings to the Research

Research Article

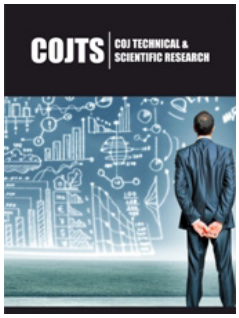
# Cyber Security Challenges and Big Data Analytics

**Roji K and Sharma G***

Department of Computer Science and Engineering, Nepal

**Abstract**

More than half of the global population is involved in the cyber world. With this massive population there is a generation of large amount of data. These data are of high-volume variety and velocity, which is collectively termed as Big data. The security of these data is of primary importance. Cybercrimes, threats are also taking the form of big data. We cannot deny the fact that we have conventional cyber security application that filters the anomalous traffic that are seamlessly working fine unto this point. But due to rapid expansion on data and threats these applications need to be updated. Big data analytics tool when incorporated with these traditional approaches may help to find the efficient measure to defeat security attacks. Here in this paper we have gone through some of the major cyber security threats, some of the conventional Cyber security application and try to find how big data analytics can be incorporated in those application to obtain the reliable outcome.




***Corresponding author:** Sharma G, Department of Computer Science and Engineering, Nepal

**Submission:** May 24, 2019

**Published:** June 18, 2019

Volume 2 - Issue 2

**How to cite this article:** Roji K, Sharma G. Cyber Security Challenges and Big Data Analytics. COJ Technical & Scientific Research.2(2). COJTS.000534.2019.



## Introduction

The internet we see today is expanding faster than we can imagine. Since the dawn of the Internet, the number of websites has gone up drastically and so has the amount of data stored on the web. Recent estimates and data released by Google [1] have revealed some very shocking facts about the Internet.

a. The Internet consisted of 5 Million terabytes of data way back in 2010.

b. As of October 2018, there are more than 1.9 billion websites on the Internet.

c. 95 million photos are uploaded on Instagram every day.

d. Internet users consumed one zettabyte bandwidth in 2016.

e. 4 Billion Out of the 7 Billion people on earth are already online.

f. 85,000+ websites are hacked every day.

g. 5000 domain names are registered every hour.

h. Facebook boasts of a massive 2.234 Billion users.

i. 400 Hours of video contents are uploaded on YouTube every minute.

These few facts present us a scenario that increasingly huge amount of important information about the users are delivered and stored on the internet. The exponential growth of the Internet interconnections has led to a significant growth of cyber-attack incidents often with disastrous and grievous consequences. According to [2], most current cyber security threats can be categorized into the following broad categories

a. Advanced Persistent Threats (APT)

b. Insider Data Theft

c. Distributed Denial of Service (DDoS)

d. Trojan Attacks

e. Phishing

f. External Software Introduction including Malware

g. SQL Injection

h. Zero-day Attack

These threats are taking the form of the massive data, which are existed in network traffic; the primary problem of traffic classification is raised by big data, which can be illustrated with three characteristics- volume, variety, and velocity [3].

I. Volume: Volume refers to the amount of data generated in any for (media, text, files) etc. the Volume here consist of files, records, transactions, archives, and logs.

II. Velocity: Velocity here refers to speed of processing of data that is being generated. The batch here consists of batch, unpredictable fast streams, real time.

III. Variety: Variety means the different and miscellaneous type of data generated. The variety here consist of structured and unstructured, semi structured etc.

Protecting network from various attack is paramount task, where traffic data analysis is a key technology [3]. While big data analytics is being leveraged for business analytics, prediction of sales and profit, and adding to business value, we can also use big data to defend against the cyber threats, prevent cyber-attacks, and improve cybersecurity and situational awareness [4]. The core or the backbone to this perspective of big analytics in cyber security is "improved detection ", and that is where big data come into play.

## Big Data Analytics

As per the definition obtained from TechTarget big data analytics is "complex process of examining large and varied data sets or big data to uncover information including hidden patterns, unknown correlations, market trends and customer preferences that can help organizations make informed business decisions." Information securing sector has been actively involving in network monitoring, maintaining system logs and managing information sources since decades. It is not clear to cyber security community about the difference between conventional approach and big data (Table 1). The summarized report of "Big Data Analytics for Security Intelligence" [5] unclouded the dissimilarities between these two approaches. Below are few of them. There are certain steps big data analytics follow for storage, analysis and maintenance [6] enumerated some of the basic procedures generally big data analytics follow. Collection of logs from many sources-In this step, the collection of data takes places from different sources. This collected data has variety of nature, some might be structured some might be unstructured. This collected data also needs to be stored for further processing. Maintaining the data consistency - In this step, the different nature data is taken and then preprocessing (the process of transforming raw data into understandable format) is done on it to make that data even. Different techniques of data preprocessing are:

i. Data Cleaning: Data cleaning is the process of filling out unavailable information, removing noise from data and deleting unwanted value.

ii. Data Integration: Data is taken and put together with different representation to resolve the conflicts within the data sets.

iii. Data Transformation: The data is normalized i.e. the redundancy from the data is removed and is generalized.

iv. Data Reduction: After the data is preprocessed it is then reduced so that it can be integrate into the data warehousing where the analytical process will take place on this data.

v. Perform correlation: In this process after the data is entered into the data warehouse the correlation and dependencies of one data set is checked with other data set to possibly find out that to which extent those sets fluctuate together.

**Table 1:** Conventional approach and Big data approach.

| Factors | Conventional Approach | Big Data Approach |
|---|---|---|
| Storage | Retaining large quantity of data is not economically feasible | Have scalable and redundant storage.eg. No SQL databases. |
| Analysis | Inefficient in performing analytics on large complex datasets. | Big data technology such as Hadoop ecosystem enabling the analysis of large-scale, heterogeneous datasets at unprecedented scales and speeds. |
| Maintenance | Management of large data -ware houses is expensive, and deployment requires strong business cases. | Hadoop framework and other big data tools are now commoditizing the deployment of large- scale, reliable clusters and therefore are enabling new opportunities to process and maintain the data. |

Applying various intelligent algorithms and analytics: Now after following above procedures, the data becomes ready to be analyzed. Different types of algorithms and analytics are applied regarding the nature of that data. This analysis will give us the insights of the data, which we would have noticed by just observation, and those, are the insights, which enables us to find out the information and also possibly find out how to treat them. Some example of various algorithms are:

i. Apriority Algorithm

ii. Naive Bayes Classifier Algorithm

Finding out patterns or behavior anomaly to detect any threat or harm Applying the algorithms to data sets produces Visualizations of the data in form of box plot, graphs, and scatter-plots etc. which makes it easy to find out any unusual behavior or any different pattern in the data sets just by observation. But various methods can also be used to find out and analyze pattern or behavioral anomaly:

a. Clustering: Clustering is known as the grouping or integration of data in such a manner that any pattern or behavior is visible easily.

b. Classification: Classification model is also used in some cases to find out pattern and connections, classification normally consist of two parts: The variable that are used. A rule to combine the values of these behaviors in order to obtain a predicted value of a given annotation.

### Big data analytics and cyber security

Cybersecurity application relies on the deep understanding of the network traffic. Intrusion Detection System, Malware Analysis and Botnet detection are some of the popular one. Incorporating big data analytics in the conventional approach of traffic analytics of these application will help to tackle the three key challenges of traffic analysis real- time classification, unknown traffic classification and the efficiency of automated classification [3]. In the paper [5] write about the advancement of big data analytics. In same paper it is presented how big data analytics tools with combination of conventional one can be effective.

### Intrusion detection system

[3] states that" An intrusion detection system (IDS) aims at recognizing malicious traffic from normal traffic. To achieve this goal, IDS scans current traffic before rerouting. Because of the heterogeneous nature of the attacks, the malicious traffic is often embedded in botnet traffic, DoS attack traffic, spam traffic, and so on." Here traffic data are collected from various sources and managing them became a challenging task. SIEM (Security and Information Event Management) tools are the conventional tools used to aggregate and correlate all the network activities and send the information to the dashboard of security analyst. Now big data tools are improving the information available to security analysts by correlating, consolidating, and contextualizing even more diverse data sources for longer periods of time [5]. SIEM is inefficient to handle large scale, heterogeneous and rapidly production of network data. This system classified the data on the basis of port and payload. If these data are classified with Statistical approach and proper machine learning algorithm is applied, then we can achieve this. IDS is an effective Cyber security application and if we use proper traffic classification, we can enhance its capabilities, thus more security to the large data.

### Malware analysis

Existing system for malware detection is mainly signature based. The detection rate of zero-day and polymorphic malware is 25% to 50% [6]. Signature based system use already defined pattern to detect the malware. They fail when malware morphs programmatically while exhibiting the same functionalities. It is not possible to have unlimited amount of signature and no databases can store them. Big data analytics is good at analyzing real time data and take adequate decision. In the paper titled "Big data: Deep learning detecting Malware", [7] they suggest the robust approach to detect malware using deep learning. Through series of experiment and analysis on the vague cases they try to prove that big data analytics is also effective in malware analysis.

### Advance persistent threat (APT)

APT is a low and slow mode attack. It uses advanced technology to monitor and track the important information from the network without owners, consent. Here to overcome this threat analysis of wide range of data is necessary. Traditional tools are inefficient to provide correlation and pattern of this type of threat. The research presented on "The Analysis of the APT Prelude by Big data Analytics" [8] shows that using big data analytics and machine learning receive effective response to deal with APT. The same report also claims that there were no such tools to detect such anomaly up to that point of point. These are few issues currently hitting the security of digital world and as aforementioned they are of various types and nature. Here going through various literatures, it come upfront that big data analytics can be the best weapon to get over cyber security attacks. Challenges: Though it is one of the promising approaches. It still has some challenges. The literatures show following challenges:

**Privacy:** Big data analytics tools are effective in analyzing and correlating large set of data. So, we cannot deny the fact that these sets may contain potentially sensitive dataset. The information related to such as defense policy of a country, organizational strategy needs to be confidential. So, while correlating and analyzing such type of data proper guidelines need to be made. So that on the basis of that privacy can be maintained. If this is not done whatever approach we follow to secure the system will go in vain. So, the matter of secludeness must be considered.

**Provenance:** Big data analytics never bother about the origin of data. Here in some literature framework are suggested, various sample of data are taken. Big data analytics of data consists of large scale of data. The authenticity and integrity of the data are very critical. It is hard to find the trustworthiness of data that are considered in those researches. Meanwhile, sources of data are not made transparent to reader on one side and in other that are not given so much importance. Human Computer Interaction: Statistical big data analytic tools are on action but not about visual analytics. We can focus on this side of big data analytics. Incorporating human computer interaction in big data analytics can surely bring unimaginable outcome in cyber security industry. In other hand it will make big data tool more popular then now.

## Conclusion

We are in the stage where cyber security cannot be compromised. There exist conventional tools to fight these issues, but they are not sufficient. We need to blend traditional tools with big data analytics tools to acquire better and more secure system.

## References

1. https://lifehacks.io/facts-about-the-internet/
2. Eastman R (2015) Big data and predictive analytics: On cybersecurity frontline. IDC Whitepaper.
3. Miao Y, Ruan Z, Pan L, Wang Y, Jhang J, et al. (2018) Automated big traffic analytics for cyber security. IEEE 1(Xiv:1804:09023): 1-5.
4. Joglekar P, Pise N (2016) Solving cyber security challenges using big data. International Journal of Computer Application 154(4): 9-12.
5. Cardenas A, Manadhata P, Rajan S (2013) Big data analytics for security. IEEE Security & Privacy 11(6): 74-76.
6. Apurva A, Ranakoti P, Yadav S, Tomer S, Roy NR (2017) Redefining cyber security with big data analytics. International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), Gurgaon, India, pp. 199-203.

7. Emmanuel M, Kyanda SK, Julianne SO (2018) Big data: deep learning for detecting malware. ACM/IEEE Symposium on Software Engineering in Africa. Uganda.

8. Young C, Park WD (2016) The analysis of the APT prelude by big data analytics. Journal of the Korea Institute of Information and Communication Engineering 20(6): 1129-1135.