The Amalgamation of Big Data, Machine Learning and Cyber Security

Harshita Puri
Dept. Of Computer Science HMR
Institute of Technology and Management,
Hamidpur, Delhi India

Bhuvan Sharma
Dept. Of Computer Science HMR
Institute of Technology and Management,
Hamidpur, Delhi India

Vaishali
Dept. Of Computer Science
HMR Institute of Technology and Management,
Hamidpur, Delhi India

***Abstract* - This paper contributes to a new system, which is capable of using Big Data to strengthen Cyber Security.It fills a methodological void between three domains: Big Data, Machine Learning and Cyber Security by introducing a new system. The aim is to protect every system from a data leak by analyzing current data sets and using history of attacks, ifany. The paper emphasizes on analysis of attacks and live streaming data to gather footprints of attacker. Since, data is growing in both size and multitude; hence there arises a need to protect that data from unauthorized access. Its time, a system should be intelligent enough to do the analysis and yield useful results. The ultimate goal is toautomate the method of Cyber Crime Detection.**

Keywords- Big Data, Machine Learning, Cyber Security, APTs

I. INTRODUCTION

Ever since the humans evolved, history is being created. We have been preserving that history in the form of data. From storing them on clay tablets to storing them in intelligent systems, we have been working on them. As the days passed by, data grew in size as well as multitude.The problem was that we had lots of data and we didn't know what to do with it! But Data Analysis has found its place in the market and its importance has been growing since the last decade. With faster systems, which can process efficiently, proper utilization and analysis of Big Data has become possible.There is no doubt over its vast use as various giants have already been using it. For example, LinkedIn uses Big Data to generate recommendations for over billion users. In addition, many governments have already invested in BigData, for instance in March, 2012, the U.S. Government invested 200 million dollar in BigData Projects.

Also, as mentioned by Hong-Mei Chen, Rick Kazman and Serge Haziyev in "Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach" [1] : "Web-based systems (WBS), ranging from product recommender systems, ecommerce platforms, social networking, gambling, gaming, to CRM (Customer Relationship Management), and SCM (Supply Chain Management) applications, have traditionally relied on data analytics to operate. For WBS, data analytics is about generating predictions and actionable insights to improve real-time customer experience, increase market and customer intelligence, predicts customer behaviors, optimize operational efficiency, personalize service provision, 0prevent security threats and frauds, minimize brand risk, and innovate processes and service."

Companies such as Google and Microsoft have been analyzing massive volume of data fordecisions related to business, investments and technology.

Section I contains the overview of how and why Big Data is becoming popular, i.e., how Big Data is finding its place in every domain.Section II describes the basic Big Data terminologies.
Section III defines how Big Data Analysis is carried out.Section IV describes machine learning basics along with why there is a need to include Machine Learning with Big Data in Cyber Security.Section Vemphasizes over the need of an efficient Big Data network.It also describes the characteristics of a Big Data network.Section VI describes the security aspects, and discusses about the types of threats which can be dealt by Big Data.Section VII describes the solution to Cyber attacks using Big Data. SectionVIII sums up the whole paper, i.e., the conclusion.

II. BIG DATA

"Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capturing, storage, analysis, datacuration, search, sharing, transfer, visualization, querying, updating and information privacy" as on 7th April 2017 by Wikipedia.

Big Data can be described by five V's –

- VOLUME- The amount of data is increasing day by day and hence the volume of the data is increasing.

- VERACITY- Refers to quality, messiness and accuracy of data.
- VARIETY- Data is generated by every field, every action of humans and therefore there exists sufficient variety of Data.
- VELOCITY- The speed at which data is increasing is enormous and hence the velocity of data is high.
- VALUE-The result decides the value of Big Data. If excellent results have been derived from the research, then it can be said that value has been added to it.

III. BIG DATA ANALYSIS

As stated in "The Role of Data Science in Web Science" by Christopher Phethean, Elena Simperl, ThanassisTiropanis, RamineTinati, and Wendy Hall, University of Southampton,[2] "The rate at which data is created leads to challenges in storing, managing, and using the data productively to classify a training dataset and develop a predictive algorithm to automatically classify future cases. Techniques include data mining to identify patterns and extract key relationships from the data variables and machine learning to improve predictive algorithms".

Hence, the more the data, the more is the processing required, the efficient the systems need to be and for that Big Data Analysis needs to be carried out in every domain. These tasks are time consuming as well as brain consuming!

Some Big Data analysis techniques are:

- Association rule learning- Works on the rule of Correlation. By relating things together to obtain outcomes.
- Classification tree analysis- Works on historical data, is a kind of statistical Classification
- Genetic algorithms- Inspired by the evolution mechanisms, employ optimisation.
- Machine learning- Makes predictions by learning from sets of data.
- Regression analysis- Works on dependent and independent variables. It works well with quantitative data.
- Sentiment analysis- As the name suggests, it researches on the sentiments and obtains sentiment-based results i.e. certain number of target audiences' sentiments are recorded based on a specific situation and relevant results are derived.
- Social network analysis- As social networking is on the go, therefore this holds an important place in the business market.

Some Big Data analysis tools are:

- Hadoop - Promoted by yahoo presently.
- Pig - Developed at Yahoo, for working with big data
- Hive -A SQL-like query language for big data in a warehouse configuration.
- HBase -Used as a datastore, MapReduce jobs can be executed on this.
- BashReduce - Implements MapReduce for standard commands of Unix
- Mahout - library for scalable machine learning, a part of which can use Hadoop.

Others are Disco Project, ZooKeeper and Chukwa etc.

IV. MACHINE LEARNING

Learning provides a set of algorithms and techniques, mostly of them based in statistics and probability, inferring an approximate model when provided enough observations of the target system. Learning from these massive data is expected to bring significant opportunities and transformative potential for various sectors.

Deep learning or majorly Machine Learning and BigData are the buzzwords of the industry as well as in the recent advances of cyber security. Learning as one may consider is the ability to acquire knowledge and to use it to produce results. The Learning Process: Measurement, Programming, feature selection or feature projection, model learning analysis results is of utter importance in machine learning. As the world is progressing towards a better technology, data has been increasing for obvious reasons. New malwares are finding their place in the market; zero day vulnerability and data leaks have become very common. Our aim is to remove the data leaks from the experts' end to the thieves' end. As more and more companies or almost every domain uses Big Data, with vast multitude of data sets, those data sets are now at risk or have been at risk always but have never been paid any heed.

V. BUILDING A BIG DATA NETWORK

As stated in- Big Data Analytics: Security and Privacy Challenges by Youssef Gahi, MouhcineGuennoun, Hussein T. Mouftah [3], "There is a set of privacy and security concerns that must be considered before building a Big Data environment. In what follows, we highlight the most important challenges that should be taken into consideration when dealing with Big Data.

1. Random Distribution 2. Privacy 3. Computation4. Integrity 5. Communication"**,** the challenge before us is to build a network with BigData which is secure enough to deal with attackers.

The process is somewhat like:

We have a HDFS along with data node and name node.The workload execution would be as follows:

- Map Phase- The input is split into number of phases. The mapper processes the input split, one at a time.
- Shuffle phase- A handoff phase, performs sorting.
- Reduce Phase-Combines the data to obtain useful results.

Then, the data is read by the application.The place where the bits hit, is the place where map reduce is employed .In BigData, we need networks that receive big chunks of data suddenly that is described as a push and after that suddenly the bandwidth might go down and everything is quiet! So, a network which is somewhat like this and supports it wholly is required with BigData.The characteristics therefore required for BigData networks are:

- Availability: The performance is affected if a loss of portion of data occurred.For instance if you have 50 data nodes, then you could be losing around 15 data nodes! Hence your capability to process faster and in an efficient manner goes down.
- Burst Handling: Queue depth and low over subscription ratio is important.
- Jitter: The variation in delay is jitter.Even a delay of 5 micro seconds is a delay.
- Latency: It should be low.
- Security: Our resources and data should be free of any kind of unwanted intervention.

Solution to the problem of building a network: What is required is scaling up of environment quickly because we deal with petabytes and zetabytes of data.

VI. THE SECURITY ASPECTS

"In general, Information Security is a domain problem, not a domain solution and hence it seeks solutions from other fields or domains."

Traditionally, security problems were aided by mathematical methods.Some of them are:

- Secrecy- Using Cryptography i.e. the art of solving codes.
- Integrity- To keep the integrity of a system, we use Hash functions.

Coming to firewalls which are the necessity for today, we get secured by them but Big Data and Firewalls make for slower processing because the more confidential your data is, the better firewall you employ thereby slowing down your system because then IP addresses, router and many things interfere in between! But eventually we need a firewall.

A. Types of threats:

Traditional attacks: The goal is to infect a multitude of machines and to acquire their resources. These can be tracked through analysis using BigData as they can be easily detected in comparison to Advanced Persistent Threats.

APTs(AdvancedPersistentThreats)

In this, the attacker can exist in system as long as its goal is not achieved and it can cope with the traditional security and has the potential to steal the information. This is very similar to a zombie attack. The attacks are very much unique in nature with no available history hence they bypass the security mechanism thereby completing their goal without being detected! They steal bare minimum amount of resources which are below the radar level. The solution to such APT attacks is to detect their existence in past if any, and to determine their goal and then to secure our systems, we can make goal completion-prevention systems. Hence, to fight against persistent attacks, we need to collect a large amount of data. Then the next step is to analyze those attacks to identify the weak link. Apart from such methods, Machine Learning can be used to tackle security challenges like Malware detection- “a file which contains malicious software payload”.

APTs are difficult to analyze and therefore we find BigData somewhat ineffective in dealing with them. Also, it is not possible to predict that who will be the next target of such an attack; hence the problem of APT remains unsolved using BigData.

DEALING WITH SECURITY RELATED ISSUES-, there is a need to tackle the security relate issues and to deal with the privacy concerns too.

VII. BIG DATA- A SOLUTION

With BigData, we can collect the data of the person or organization attacked in earlier times and then using Big Data, an analysis can be carried out regarding the type of attack, goal of the attack and the attacker can be traced!

How Big Data can prevent security attacks is in a mechanism as follows:

The attacker attacks the system and bypasses the traditional security but if he is not a Data analyst then itis extremely difficult to understand what the huge dataset contains.

Also, APT attacks steal minimal information and in BigData that would not work because unless all the information is linked, it is of no use and carrying out the APT attack again and again would mean a long wait as 1 TB of data would take years to be extracted. The requirement therefore is to predictor act on all the security threats and then to analyze them closely. Analysis of streaming data encompasses us to protect our system in real time.

These attacks can be determined and detected by following methods:

- Identifying patterns by employing BigData on email, this will identify the number of redirects as well as targets.
- Identification of suspicious domains.
- Use of forensic analytics would help link the identities by analyzing structured attributes and the communication that are unstructured.
- Management of logs.
- Advanced threat visualization

The procedure to determine and detect these attacks is:

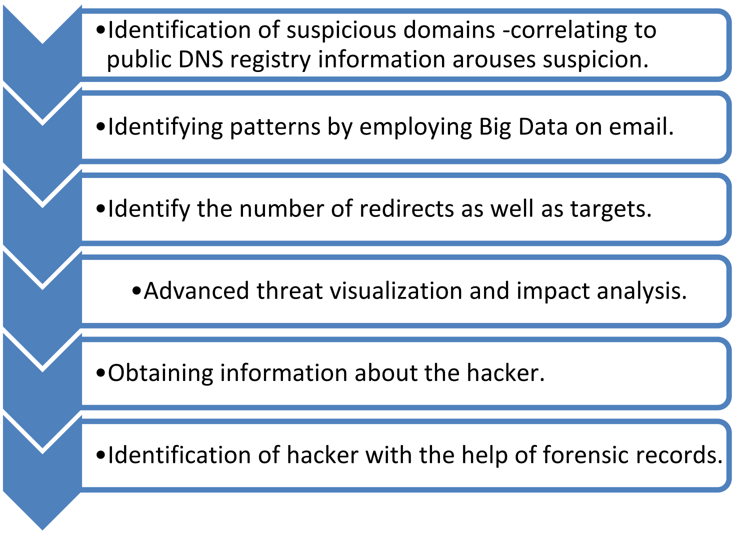
•Identification of suspicious domains -correlating to public DNS registry information arouses suspicion.
•Identifying patterns by employing Big Data on email.
•Identify the number of redirects as well as targets.
•Advanced threat visualization and impact analysis.
•Obtaining information about the hacker.
•Identification of hacker with the help of forensic records.

Fig 1: Procedure to determine and detect the attacks

What can be done is analysis of data so that a check can be kept on:

- Web Crawling- Browsing the World Wide Web in an automated manner.
- Text Analytics- Analyzing text or pattern and obtaining significant results.
- Pattern Detection- Security camera infected with a malware/s.

Later we shall realize that we need automated systems or in other words, we want the entire content to be automated that too in such a way that live streaming data can be analyzed besides being used.

VIII. CONCLUSION

Big Data plays an immense role in big decisions, be it related to business or adopting a new technology. We have been using it but not majorly in Cyber Security. With its help, live streaming data can be operated upon to detect and analyze major threats and thereby allowing us to detect those hackers. To track someone who is intervening and stealing our resources, we can employ systems which can analyze and obtain results; the only need is to develop efficient Big Data networks. With the combination of Big Data, Machine Learning and Cyber Security, we can build a system which can obtain data, analyze it and give us the result which will be the details of the person or organization which tried to intervene into our system or access our data through unfair means. So, to conclude, there is a need for automated systems or in other words, we want the entire content to be automated.

REFERENCES

[1] Hong-Mei Chen, Rick Kazman and Serge Haziyev ,"Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach", IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 3, JULY-SEPTEMBER 2016.

[2] Christopher Phethean, Elena Simperl, ThanassisTiropanis, RamineTinati, and Wendy Hall, University of Southampton, "The Role of Data Science in Web Science", IEEE Co*mputer Society, May-June 2016.

[3] Youssef Gahi, MouhcineGuennoun, Hussein T. Mouftah, "Big Data Analytics: Security and Privacy Challenges"[School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward Ave., Ottawa, ON, Canada].

[4] A.A. Cardenas, P.K. Manadhata., and S.P. Rajan, "Big Data Analytics for Security," In IEEE Security & Privacy, Vol. 11, No. 6, pp.74-76, 2013.

[5] D. Agrawal, A. El Abbadi, and S. Wang, "Secure and PrivacyPreserving Database Services in the Cloud," In Proceedings of the 29th International Conference on Data Engineering, pp. 1268-1271, 2013. [12] M. Jensen, "Challenges of Privacy Protection in Big Data Analytics," In Proceedings of the International Congress on Big Data, pp. 235-238, 2013.

[6] S. Matthew, S. Christian, H. Benjamin, and V. Gabriele, "Big Data Privacy Issues in Public Social Media", In Proceedings of the 6th IEEE International Conference on Digital Ecosystems Technologies, pp. 1-6, 2012.

[7] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. M. M. Weijters," Workflow mining: a survey of issues and approaches. Data & Knowledge Engineering", 47(2):237–267, 2003

[8] A. Cuzzocrea, I-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the Big Data revolution!,", In Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, pp. 101-104. 2011.

[9] X. Wu, X. Zhu, G-Q. Wu, and W. Ding, "Data mining with Big Data," In IEEE Transactions on Knowledge and Data Engineering, Vol.26, No.1, pp.97-107, 2014.

[10] R. Buyya, K. Ramamohanarao, C. Leckie, R.N. Calheiros, A.V. Dastjerdi, and S. Versteeg, "Big Data Analytics-Enhanced Cloud Computing: Challenges, Architectural Elements, and Future Directions," In arXiv:1510.06486, 2015.

[11] J.P. Nivash, E. Deni Raj, L.D. DhineshBabu, M. Nirmala, K.V. Manoj, "Analysis on enhancing storm to efficiently process Big Data in real time," In Proceedings of the 2014 International Conference on Computing, Communication and Networking Technologies, pp.1-5, 2014.

[12] Daniel E. O'Leary, University of Southern California," Ethics for Big Data and Analytics", IEEE Computer Society, July-August 2016.

[13] N. Marz and J. Warren, "Big Data: Principles and Best Practices of Scalable Realtime Data Systems", Greenwich, London. U.K.: Manning Publications, 2013.

[14] JosepLluisBerral-García Barcelona Supercomputing Center, Jordi Girona 29-31,"A Quick View on Current Techniques and Machine Learning Algorithms for Big Data Analytics ",ICTON 2016,We A3.1.

[15] George Strawn, US National Academies of Sciences, Engineering, and Medicine,"Data Intensive Science", IT Pro September/October 2016.

[16] Xiaoyin Sun1, Wei Zhou1, Qiyan Xiang1, Binyue Cui2, Yi Jin1 1 School of Computer and Information Technology, Beijing Jiaotong University, Beijing China 2 Hebei University of Economics and Business, Shijiazhuang China," Research on Big Data Analytics Technology of MOOC", The 11th International Conference on Computer Science & Education (ICCSE 2016) August 23-25, 2016. Nagoya University, Japan.

[17] Sam Supakkul, Liping Zhao, Lawrence Chung, "GOMA: Supporting Big Data Analytics with a Goal-Oriented Approach, 2016 IEEE International Congress on Big Data.