Solving Cyber Security Challenges using Big Data

Prajakta Joglekar
Computer Engineering Dept
MAEER's Maharashtra
Institute of Technology, Pune, India

Nitin Pise
Computer Engineering Dept
MAEER's Maharashtra
Institute of Technology, Pune, India

ABSTRACT

Cybersecurity has become a Big Data problem as the size and complexity of security related data has grown too big to be handled by traditional security tools. In this paper the authors have described the categories of cybersecurity threats and challenges posed by them. They have analyzed how big data tools and concepts are being used to solve these challenges and detect and prevent attacks real-time.

General Terms

Big Data, Security

Keywords

cybersecurity, security, big data, attacks, APT, malware, malicious, Hadoop

1. INTRODUCTION

It is the era of Big Data. We are living in the digital world today, where data is getting generated every minute of the day. According to Intel's Infographic [1], 1,572 GB of global IP data is transferred every Internet minute. 38,194 photos are loaded on Instagram, 4.1 million searches are done on Google, 194,064 apps are downloaded while 3.3 million pieces of content are shared on Facebook in an Internet minute. It is predicted that by 2017, the number of connected devices will be equal to three times the number of people on earth. This massive traffic will add to the volume and variety of Big Data. While the big data technology 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. This paper describes how big data can be used to strengthen cybersecurity. The paper first starts with the definition and concepts of cyber security, different categories of security and challenges faced in cyber security. It then describes the characteristics of Big Data and how it can be useful to strengthen cybersecurity. It also discusses few big data tools which are being used to solve cyber security issues. In general, the paper does a survey of how paradigm shift is occurring in cybersecurity due to Big Data Analytics

2. CYBERSECURITY

Wikipedia [2] defines Cybersecurity as the protection of information systems from theft or damage to the hardware, software, and to the information stored on them. It also includes protection from snooping or misdirection of the services they provide. In simple words, cybersecurity refers to the tools and practices implemented to protect information in this digital world. According to the IDC IView Report on 'Digital Universe in 2020' [3], the proportion of data in the digital world that needs protection is growing faster than the digital world itself, from less than a third in 2010 to more than 40% in 2020. Forbes noted that Target store, the second largest discount retailer in the United States, spent $1.6 million on one of the most advanced security products available, yet still missed the attack that cost a CEO and CIO their jobs. Cyber-crime costs are predicted to reach $2 trillion by 2019. Such big corporate hacks do emphasize the need to upgrade the security systems and practices.

3. TYPES OF THREATS

According to Robert Eastman [4], most current cybersecurity threats can be categorized into the following broad categories:

3.1 Advanced Persistent Threats (APT)

An APT is a set of quiet and continuous computer hacking processes, often coordinated by humans targeting a specific entity. An APT usually aims organizations or nations for business or political purposes.

3.2 Insider Data Theft

An insider threat is a malicious threat to an institute that comes from people within the institute, such as employees, contractors or business associates, who have inside information concerning the institute's security practices and data. Data theft is done with the intent to compromise privacy or gain confidential information.

3.3 Distributed Denial of Service (DDoS)

In computing, a denial-of-service (DoS) attack is an attempt to make a network resource unavailable to its intended users, by temporarily suspending services of a host connected to the Internet. In a DDoS the attack source is more than one, often thousands of unique IP addresses.

3.4 Trojan Attacks

A Trojan horse is any malicious computer program which appears as useful, routine, or interesting in order to influence a victim to install it. Trojans are commonly spread by some form of social engineering.

3.5 Phishing

Phishing is an attempt to acquire sensitive information such as usernames, passwords, and credit card details, by impersonating as a trustworthy entity.

3.6 External Software Introduction including Malware

Malware is any software used to disturb computer operations, gather sensitive information, and gain access to private computer systems. Sometimes, it can be used to display unwanted advertising.

3.7 SQL Injection

SQL injection is a code injection technique. It is used to attack data-driven applications. Malicious SQL statements are inserted into an entry field for execution. An SQL Injection can destroy your database.

3.8 Zero-day Attacks

A zero day susceptibility refers to a security hole in software that is not known to the vendor. This hole is then misused by hackers before the vendor becomes aware and tries to fix it—this exploit is called a zero day attack.

3.9 URL Redirection or Parameter Tampering

The web parameter tampering is constructed on the manipulation of parameters exchanged between client and server so as to modify application data, such as user credentials and permissions, price and quantity of items, etc. Generally, this information is stored in cookies, hidden form fields, or URL Query Strings.

The threat actors for the above categories can be classified as insider, opportunist, accidental user, hacktivist, state-level actor or professional criminal.

4. CHALLENGES IN CYBERSECURITY

Cyber-attacks are global and the risks associated with cyber security are universal. These are not the concerns of any particular country alone.

- An article from CSI [5] observes that satellites, power grids, thermal power plants, websites, banks and almost all digital systems, are susceptible to cyber-attacks. Thus, while the world is going digital and progressing by leaps and bounds, this progress requires security.
- Attackers are extremely advanced and have organized crime rings which run attack operations on a large scale. They are seen to have tremendous collaboration with each other, which is lacking among the good guys.
- The authors of [6] have proposed nine D’s of cybersecurity to detect, react and sustain against the attacks. These include deterrence, detection, differentiation of protection and distraction with decoys.
- IDC's research uncovered that cybersecurity threats are evolving at a rapid rate and that firms and government agencies must shift from a reactive approach to a pre-emptive approach[4] by understanding the threat before an attacker can cause damage.
- Shifting to a proactive approach requires organizations to look into all available information and apply predictive and behavioral analytic tools to discover the likelihood of a threat, detect the actual threat, gather intelligence about the attack, and execute an enterprise wide response before the threat becomes significant.
- The historical dump-and-then-analyse approach has proved to be ineffective because the needed data history is not typically stored or analysed in a timely fashion.

Enter Big Data!

5. BIG DATA

This section takes a look at what Big Data is, and how it can help one to solve cybersecurity challenges. According to Gartner’s definition [7], big data is high volume, high velocity and high variety data that demand cost-effective and state-of-the-art information processing to produce value for the business. It is an interdisciplinary issue which requires the collaboration of academia, industry and enterprise [8].

5.1 Sources of Big Data

- In USA, most of the government data is open data. This includes demographic as well as medical data of US citizens. Many other countries, including India are following the suit.
- Facebook is a huge source of data which users share with the real world.
- Google provides statistics on search volume, which itself is a big data.
- Financial Data Sets are available at financial data finder at OSU.
- Twitter provides live streaming data through its APIs.
- Log files of an application tend to increase in volume thereby becoming a big data source.
- IoT sensors and devices continuously generate data.

5.2 Applications of Big Data

Big Data is mainly used to understand and optimize business processes, do financial trading, improving and optimizing smart cities and nations, understanding and targeting customers for better relationship management, improving healthcare, sports, transport services etc. Big data is also used to improve cybersecurity, which is the topic of our work.

6. HOW BIG DATA CAN HELP IN CYBERSECURITY?

- Mining usable information from large amounts of data offers a broader view of risks and vulnerabilities. Big data can therefore be used to enhance security.
- New big data tools can efficiently handle the complexity and volume of IP Network data which is required to analyze cyber security.
- Various machine learning algorithms for classification and prediction make it possible to identify abnormal behavior much earlier.
- Since big data Tools can handle variety of data – structured and unstructured, finding anomalies becomes easier.
- According to one of the top 12 predictions for 2016 from leading cybersecurity experts [10], new data sources arising from the Internet of Things and biometrics will lead to a renewed government interest in using big data to prevent terrorism.
- As per the research survey conducted by TeraData Corporation [14], to make their organizations more secure, managers would prefer big data analytics collaborate with anti-virus/anti-malware, anti-DoS/DDoS, security intelligence systems (SIEM) and content aware firewalls.
- Financial Services are more aggressively moving towards implementing big data analytics for cybersecurity, as compared to government services.
- Commodity hardware and Hadoop framework provides the ease of collecting and storing large amount of data on which analytic techniques can be applied in order to identify breaches or malware.

- More the variability of data, the more are the chances of correct assessment and classification by a training model.
- On the velocity aspect, since big data tools can quickly iterate through the data, build adaptive models and provide quick visual analysis, that too using commodity hardware, it makes things easier for a cyber security analyst who has to analyze millions of records every day.
- From a cybersecurity point of view, data models need to have predictive power to automatically differentiate between normal network traffic and abnormal, possibly malicious traffic that can indicate an active cyberattack or malware infection.

7. BIG DATA TOOLS FOR CYBERSECURITY

7.1 Apache Spark

Apache Spark is a fast engine for data processing on a large scale. It is an open source cluster computing framework. According to the authors of [13], Apache Spark can help cybersecurity officers analyze data and answer questions:

- Which internal servers of the company are trying to connect to internationally based servers?
- Has user's access pattern to internal resources changed over time?
- Which users exhibit irregular patterns of behavior such as connecting using non-standard ports?

Spark powered big data discovery solutions can be used to detect anomalies and outliers within large datasets. Visualization techniques help when petabytes of data is to be analyzed.

7.2 Fortscale

Services Fortscale is a big data solution against APT attacks [15]. APT attacks can take place over a stretched period of time while the victim organization remains ignorant about the invasion. According to Fortscale, big data analysis is a suitable approach for APT detection.

- A challenge in detecting APT is the massive amount of data to examine through in search of abnormalities.
- The data comes from an ever-increasing number of miscellaneous information sources that have to be audited.
- Fortscale uses Cloudera Hadoop distribution to address big data challenges, and examine network traffic data to check for invasions if any.
- Fortscale employs data science techniques like machine learning and statistical analysis to adapt to changes in the security environment.

7.3 IBM Security QRadar

This tool uses big data capabilities to help keep pace with advanced threats and prevent attacks proactively [16]. It helps reveal hidden relationships within large amounts of security data, using analytics to reduce billions of security events to a controllable set of prioritized incidents. It uses the following features of Big Data solution:

- Real-time correlation and anomaly detection of security data, which is diverse in nature.
- High-speed querying of security intelligence data.
- Flexible big data analytics across structured as well as unstructured data – this includes security data, email, document and social media content, business process data; and other information.
- Graphical front-end tool for visualizing as well as exploring big data.

8. BIG DATA PROJECTS RELATED TO CYBERSECURITY

- Platfora is Big Data Analytics platform built on Apache Hadoop and Spark. Platfora provides platform for security event pattern processing to identify malicious activity [17].
- Companies like Niara are currently developing Cyber security tools on Hadoop Cluster. Niara's product will combine various heuristic techniques with new machine learning technologies to enable companies detect cybercriminals proactively.
- Project Metron (Apache) – It ingests security telemetry data at high speed and then pushes it to computation and analytics. The interfaces also presents alert summaries.

9. CONCLUSION

As compared with traditional cybersecurity methods and efforts, big data and behavioral analytics offer the opportunity to improve situational awareness as well as information security. In the financial services domain, Visa, MasterCard, and American Express have used analytics to detect potentially false transactions based upon patterns and pattern recognition across millions of transactions. Thus the volume, velocity and variety of Big Data can be used to solve cybersecurity issues efficiently and proactively.

The authors would further like to study these big data tools in detail and compare them on their ability to accurately predict a cyber-attack. The idea of using Big Data for solving cybersecurity challenges can be used in Smart Cities scenario to build a strong, resilient framework for the city. Urban computing and Smart Cities concept is an emerging area of research, which contributes to big data and is vulnerable to attacks. Thus, as a future scope, it would be worthwhile to build an urban framework which incorporates security issues by design, using big data architecture.

10. REFERENCES

[1] Intel Corporation, "What happens in an Internet Minute?," Infographic, year 2014, website http://www.intel.co.uk/content /www/uk/en/ communications/internet-minute-infographic.html

[2] Wikipedia, "Computer security", Wikipedia, the free encyclopedia.

[3] John Gantz, David Reinsel, "Digital Universe in 2020," IDC IView Report, December 2012.

[4] Robert Eastman, "Big Data and Predictive Analytics: On the Cybersecurity Front Line", IDC Whitepaper, February 2015

[5] N.J.Rao, "Cybersecurity: Issues and Challenges," CSI Communications, Volume no 39, May 2015.

[6] Wilson, Kiy, "Some fundamental cybersecurity concepts", IEEE access, 2013

[7] Gartner, “Big Data”, IT glossary, www.gartner.com

[8] Shen Yin, Okyay Kaynak, “Big Data for Modern Industry: Challenges and Trends”, Vol. 103, No. 2, February 2015, Proceedings of the IEEE

[9] Stephen Kaisler et.al, “Big Data: Issues and Challenges Moving Forward”, IEEE Computer Society Intl Conf in Hawaii -Jun13

[10] IBM, “Top 12 predictions for 2016 from leading cybersecurity experts”, http://www.ibmbigdatahub.com

[11] Morel, “Artificial intelligence a key to the future of cybersecurity”, ACM, 2011

[12] Zuech, “Intrusion detection and Big Heterogeneous Data: a Survey”, Journal of Big Data, Springer, 2015

[13] IBM, “Spark takes on the big security threats”, article, http://www.ibmbigdatahub.com

[14] Teradata, “Big Data Analytics in Cyberdefence”, Research report by Teradata, February, 2013

[15] Fortscale tool, https://fortscale.com

[16] IBM, “IBM Security Intelligence with big Data”, article, http://www-03.ibm.com/security/solution/intelligence-big-data/

[17] Platfora, “Platfora big data discovery platform”, www.platfora.com

[18] Gersh, Bos, “Cognitive and organizational challenges of big data in cyber defence”, ACM human-centered big data research, 2014