Information Assurance via Big Data Security Analytics

[1]Abhishek Gupta, [2]Mahesh Pawar, [3]Dr. Sachin Goyal, [4]Ratish Agrawal
[1]*Research Scholar RGNCLC NLIU, Bhopal*
[234]*Asst. Professor, UIT, RGPV, Bhopal*

Abstract – *The research presented in this paper represents the importance and leveraging factor of Big Data Security Analytics (BDSA) through models that will augment the predectiveness and possible mitigation of Advanced Persistent Threats (APT). Big Data Security Analytics is highly scalable & it can be achieved by breaking down the silos of data both structured and unstructured to find anomaly. Information security is evolving continuously so scalability is also becoming mandatory which can only be achieved by the integration of security intelligence platform and big data platform as described in the research paper. In order to cope up with APT's, Big Data Security & Control Framework (BDSCF) is proposed and described in the paper which consist Define, Dissect & Defend of huge datasets. The Big Data Security Analytics techniques, challenges and possible outcomes are presented.*

Keywords: *Advance Persistent Threat, Big Data, Big Data Security Analytics, Classifier, Security Intelligence*

I.INTRODUCTION

Various definitions of Big Data are present, most of them highlights the importance of 3V's (Volume, Velocity, Variety) but now it has extended to 5V's including Veracity and Value. Big Data is an umbrella term given to the humongous data generated continuously through global population. This type of data is capable enough to give clear insights which are helpful to detect fraud and perform predictive analysis for cyber attacks. All the data that is continuously generated through satellites, hand held devices, social media, blogs, log files, sensors etc carries potential for good business governance & decision making comprised with business intelligence. In the past information security was achieved through mathematical computation and cryptography but with advance technology revolution attacks have become sophisticated, mitigating these attacks requires a strategic approach whilst Big Data is a fundamental approach kind of building block to the analyst. Running analytics engine over the silos of Big Data generates patterns and models which could establish relationship between the events and adversaries. Thinking all about thinking applying methods other than traditional ones could help to find a better way to curb new threats. At organizational level it is not so possible to reconstruct security architecture now and then rapidly but the insights form Big Data Analytics helps to set limits in multivariate environment.

A. Big Data Analytics Ecosystem

Big Data in itself is raw item which needs to be washed dried cut and cooked appropriately wit ingredients so that flavors could come out and without proper analytics tool, harnessing the benefits is quite impossible. The ingredients involved in Big Data analytics are:

Big Data – Data with huge volume having terabytes of data or even petabytes or zeta bytes of data (structured & unstructured) that contains log files, transaction details, companies documents, files comes under the category of Big Data as filtering it will give important and hidden relationships & insights. As most of the forensic data resides within the organization the anomaly detection could be done easily and threats could be identified and classified.

Hadoop – It is open source software by Apache that runs over the Big Data and performs analytics; it collocates storage with processing for performance and replicates data for availability & reliability. Finally it is deployed on commodity hardware.

1. **HDFS** (Hadoop distributed file system) – it is a distributed file system which allocates petabytes of data to different nodes and its fault tolerant capability keeps it functioning. The functionality of HDFS is core product of the whole system.
2. **MapReduce** – a processing layer which is combination of two process viz Map which distributes the data into chunks and transfers it to nodes for preprocessing and Reduce set again gets the information from the data integrate it & gives the results as output.

Hadoop initially was launched as Hadoop 1.0 which didn't had the added layer of (YARN) yet another resource negotiator.

NoSQL – it is a model/store data in non relational way that is designed to handle web scale application. It requires availability, scalability & complex analytics.

B. **Information Assurance, Security & Intelligence**

Information Security was prior achieved through application of algorithms and mathematics but it didn't provide state of the art security now neither prevents advance persistent threats. Information assurance is the super set of information security that provides broader spectrum of information management with involvement of risk management and compliances with BCP & DR. Big Data analytics gives insight both for investigating and setting up new challenges based on past record. Every organization have some or significant difficulties to set up security data analytics. Security is incredibly important to organizations because whenever organization discovers that they have been breached there reputation and brand image is badly destroyed. The steps organization could take to achieve information assurance are:

a) Combating advance persistent threats.
b) Mitigating exposure to cyber attacks
c) Mitigating fraud on business process
d) Preventing Hacktivism
e) Identifying insider threats

Traditionally all the above issues were handled by managed by monitoring the network, databases information, log in information from the host, firewall, identity access management etc but due to the dynamic nature of malwares it's high time for intelligence to come into the picture.

all data

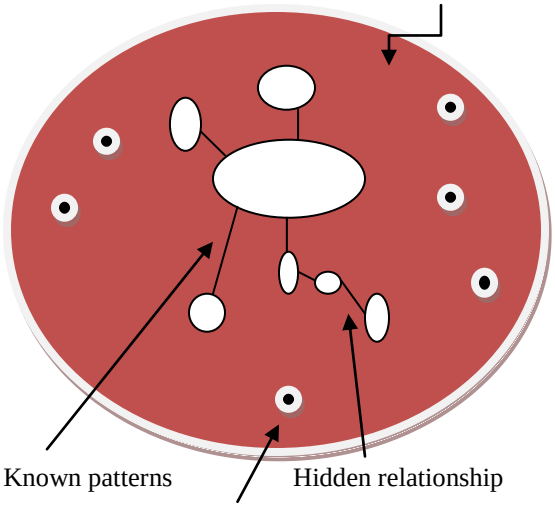


Fig1. Security data analytics

C. Adversaries, prediction & detection

Traditional attacks using malware – earlier the attacks were done to achieve one simple goal i.e. to infect as many computer as possible and create an army also knows as computer zombies/ botnets. Creating botnets helps the attacker to launch re attack and steal sensitive data or disseminate spams and cause DOS attack. Classifiers in which all the signatures of malwares were stored were able to detect these attacks but after the introduction of APT's it became impossible because the attacker spends a lot of time to design such attacks in their sophisticated labs based on zero day vulnerabilities.

Advance persistent threat – APT's are among the most sophisticated attack so far as they have their own capability and targeted goal, the biggest challenge in Big Data Security Analytics is monitoring capability which is highly compromised by APT's as they bypass it easily, for example: Stuxnet case, IRAN.

Extracting the features of malicious code is one good remedy to detect anomaly in future. The codes are generally in bit sets that could be collected and dumped into a repository & use machine learning techniques to detect malware (computer program that are capable to disrupt computer, computer network & computer resources).

Advanced – the attack is capable of bypassing present traditional security system & in many cases it is based

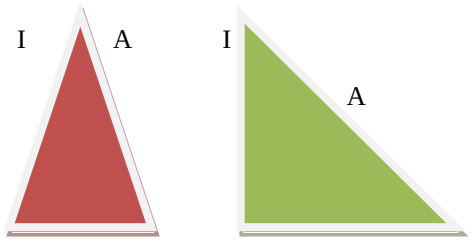
on zero day vulnerabilities. Usually APT's go unnoticed.

Persistent – advance attacks have specific goal not the whole system as it was in traditional attacks, unless the goal is disrupted the APT's keep running at very low volume and removes itself from the network after the job is completed.

Threat – APT's are specially designed to disrupt and destroy the systems by collecting & stealing information making victims system unavailable or modifying the data of victim. All these acts will break the CIA (confidentiality, availability & integrity) triad.

Characteristics of APT:

- It works at low volume to remain undetected.
- Attack is always unique in itself
- Prepared through advanced research and efforts, stays in the system for longer time.
- Erase itself without any trace after the attack.
- Usually unable to get detected by traditional tools because it is based upon zero day vulnerabilities.



Traditional ---------- > APT's

Fig2. Advanced threat impact.

Traditional methodology is lacks in mitigating threats arising from web, file systems and malware in all together. Firewalls, IDS, IPS are not capable of detecting APT Eg: malicious code is attached in the PDF is sent to the VP of an organization that unknowingly opens it, the malware existed in the resource for long time and gathered all the sensitive information and transferred it to the source (attacker) then after completion it removed itself from the network making it theoretically invisible. In order to cope up with such attacks like APT's there is a need to collect large amount of data from different source and break down it by adding context & filters using advanced analytics techniques to detect the attack.

II. BIG DATA SECURITY ANALYTICS

As more and more integration of ICT is taking place at small and large scale organizations the dependency on these system is have also increased making organizations more vulnerable and open to cyber attacks (theft, disruption, destruction). From an analyst lens:

a) It takes many days to know that the organization is breached.
b) Few of the system are infected as botnets.
c) During investigation it is found that always valid credentials were used.
d) Majority times the third party discovers the breach.
e) APT compromises monitoring & bypasses the detection system.

For identifying threats organizations has to think beyond traditional scanning, monitoring and detection tools & need to identify & protect against threats by building broader insights from broader datasets. Organization have to look into Big Data & incorporate five golden rules i.e**. acquire, organize, analyze, automate & integrate** Big Data to overcome the challenges.

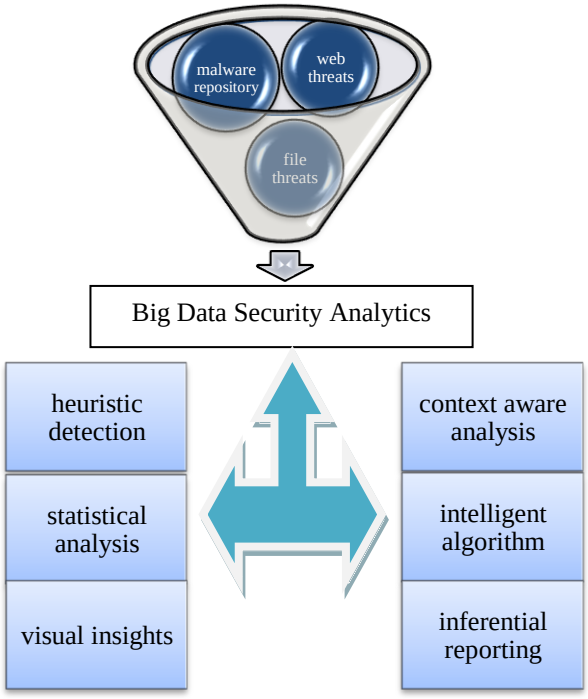


Fig.3 BDSA features

A. Integration of Security intelligence & Big Data platforms.

Intelligence plays a key role in business decision making and detecting attacks plus helps in performing Digital Forensics. The aggregation of Big Data Analytics and security intelligence gives clear insights that are capable enough to draw a security control framework as per the organizations demand.

Security intelligence platform – it is structured, analytical & repeatable. This platform contains two processes viz:

- Relative Processing is real time networking data correlation with anomaly detection. It also posses event and flow normalization in distributed architecture complimenting security context & enrichment.
- Security Operation is a workflow management is maintained with predefined rules & compliance reporting creates activity & event graphing within the organization. It also performs offense scoring & prioritization.

Big Data platform – it is exploratory, intuitive in nature with lot of creativity.

- Big Data processing is preservation of raw data (structured & unstructured) for long term in multi storages. It also features enterprise integration with real time stream computation in a distributed Hadoop infrastructure.
- Analytics & Forensics is the key feature providing predictive & decision model that creates pictographical results. It is also capable of resolving ad hoc queries; it contains collaborative sharing tools with intuitive user interface that gives interactive visualizations and help in understanding insights.

All these features make BDSA a great component to combat APT's & deploy good governance. Information assurance is not guaranteed by BDSA but it is the best at present. Security intelligence is capable of delivering insights only due to BDSA and it also supports scalability. Harnessing BDSA for information assurance can only be achieved if data science is brought into the main picture. Usually organizations are much focused on monitoring log data, transaction records, business process data etc but analyzing and detection part is missed.

Technical challenges

BDSA is a fundamental approach and building block of the information security domain, like cloud computing BDSA will also take time to establish and smaller organization has a view that they don't have enough data to feed BDSA and get motivating outputs. Still BDSA is in its early stages and carries the potential to harness raw data to give important insights.

1. Collection of high volume networks and DNS events.
2. Cope up with the rapidly changing identifications
3. Identifying subtle indicators and
4. Integration of external intelligence
5. Collection & processing of unstructured data (social media, blogs, emails, satellites etc)
6. Integration with surveillance platform activity workflow
7. Text/voice linguistic & identity analysis with monitoring of social media & mobile content.
8. Integration with intelligence platform.

Regulatory challenges

In India no such rules, regulations or standards have been published or mandated which facilitates or challenges the BDSA in private or public organizations but it is implicit that every technology is a double edged sword. There are various legal issues viz:

1. Licensing (collection, accessing, dissemination, modification, processing) issues are not defined yet.
2. Intellectual property issues are a big concern
3. Absence of legal framework ensuring secure data storage, transactional logs & granular audits.
4. Electronic evidence & other litigation issues
5. Normative setup & compliance issues of big data security control framework.
6. Security & privacy aspects of Big Data & analytics.

Data management & limit of access in terms of having an appropriate legal framework for protecting secure computations in distributed programming framework have become a mandatory requirement. The functionality of BDSA is unmatched in providing insights and intelligence to prevent cyber threats but without backing of proper legal framework which ensures sustainability and standards BDSA incorporation will eventually lag in proper implementation.

III. PAST WORK & FUTURE MITIGATION STRATEGIES

The sudden burst of data had made organization to think upon Big Data and utilizing it rather than just storing it in huge silos. Earlier security information &

event management (SIEM) or security information management (SIM) use to take care of the data which was collected from the external source. In 2000 IDS was introduced which changed threat management scenario and till 2004 IDS started experiencing downfall because PCIDSS (Payment card industry data security standard) came into picture and gave compliance reporting a big hype. Till 2011 the management started using metrics & that was the initiation of security analytics which is now moving to BDSA.

SIM use to aggregate information regarding alarms, IPS, IDS, firewall and provided to security analysts. It used tree level model that used to correlate high level incidents like Hacktivism, intrusion, breach etc with simple events like antivirus signatures and firewalls etc. the scope of SIM was limited & aptitude to handle huge data sets was very low. Security analytics applied with security intelligence is the future of SIM. As of now the volume of Big Data is unattractive to attackers, but organizations are still using traditional tools to mitigate advance attacks (APT's), some of tools applied so far at organizational level: IDS, firewall, SIEM, IPS, Antivirus, Password Policies, metrics, log management etc.

Future mitigation strategies

As forensic data usually resides in the organization and it is theoretically impossible to remove all the trails combating APT's is possible via BDSA. The goal in relation to APT's is to detect the existence of the attack in the past, detect the goal of the attack & finally detect the source of the attack. This all could be easily done by running a separate reporting & presentation engine over the IT Big Data storage and adding a abstraction layer of Big Data Security Analytics in between these two layers.

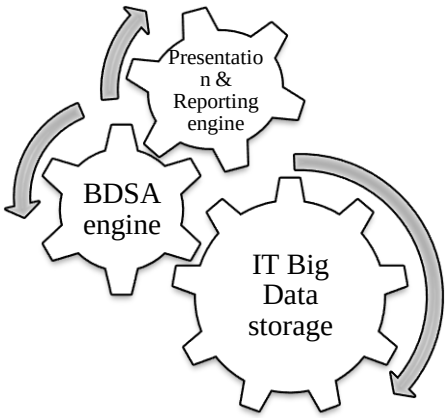


Fig4. BDSA layered structure

Big Data Security Analytics Control Framework

The research in this paper describes the functionality, aggregation and relevancy between information assurance and Big Data Security Analytics. The requirement of organizations to have spectrum of predective analytics & pattern making through advanced analytics is also described, but now at the organization level it is very difficult to implement such tools totally and effectively. For this a framework has been propsed which focuses on the integration of security intelligence with digital forensics and BDSA.

This model represents the process that should be followed to get optimum yield of applied

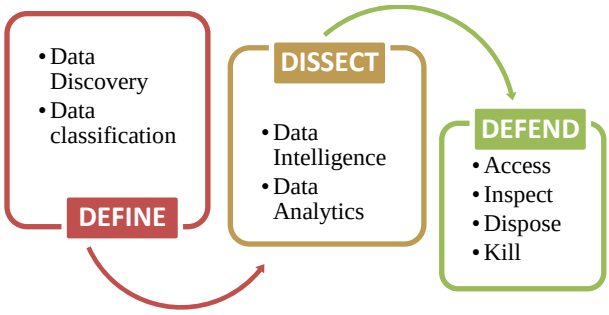


Fig5. Big Data Security Analytics Control Framework

intelligence and analytics for classifying data into benign data and suspicious data. The framework pushes to first define the data, where it is & where its classification. Next the data needs to be dissected & have some intelligence around it & provide advance analytics so that the most important ingredient i.e. context could be generated, it has to be yielded so that data could be defended. To a piece of data only four things could be done viz:

1. Putting access control to it so that checks & balances could be done
2. Inspect it for network/web traffic to find anomaly and insights
3. Delete it or disposing off the data
4. Killing it through encryption, data masking or applying tokenization.

These all four methods are like meters which need to be regulated by the analyst so that balance could be maintained, obviously all the meters could not be kept at high that's practically infeasible.

Other technologies: usage of data exfiltration software that alarms whenever data leaves the organization. Machine learning which works on artificial intelligence can also be integrated with Big Data and

enhance predictive analytics and can effectively protect the organization.

Does BDSA really work?

BDSA works & it is important because attacker doesn't know which type of data is being stored by the organization & over which data the analytics engine is running. Analytics methods are dynamic and keep on changing so this causes friction for the attacker. Some of the forensic data/evidence cannot be removed from the system. Behavioral analysis helps to predict next step and makes proactive for any intrusion. Threat happens at real time & analytics has capability to detect and analyze in real time though it requires high sophistication. BDSA is science not method so it is prominent that vast amount of research and development is still in the pipe line.

IV. Conclusion

APT's are inevitable & they will breach either for theft, disruption or just lying in the system potentially containing the attack. Attackers are hidden army not single person, no organization can guaranty that they haven't been breached it's all about when it happened. BDSA analyze history to give future in security, following the Define, Dissect & Defend model will enhance data identification & establish connection between events (noticed/unnoticed) to do predictive analysis & advance forensics. Advance analytics aggregated with machine learning, data exfiltration software will give robust protection to organization & theirs IP. Motive & source of the attack is imperative because it helps to understand why & how, also it describes the capability of present measures. Having a predictive analysis by applying Big Data Security Analytics to large variety of datasets obviously helps to protect organizations assets and customer.

V. References

[1] Nina Godbole, Sunit Belapure 'Cyber Security, understanding Cyber Crimes, Computer Forensics & Legal Perspectives' Choudhary Press, New Delhi, 1st Edition, ISBN: 978-81-265-2179-1.

[2] Nina Godbole 'Information systems security' Security management, Metrics, Frameworks & best practices, Wiley India Pvt.ltd 2013, ISBN: 978-81-265-1692-6.

[3] William Hurst, Madjid Merabti, Paul Fergus 'Big Data Analysis Techniques for Cyber-Threat Detection in Critical Infrastructures' 2014 28th International Conference on Advanced Information Networking and Applications Workshops
2014, 978-1-4799-2652-7, DOI 10.1109/WAINA.2014.141.

[4] Rasim Alguliyev, Yadigar Imamverdiyev 'Big Data: Big Promises for Information Security' IEEE

[5] Michele Chambers, Michael Minelli, Ambiga Dhiraj, 'Big Data, Big Analytics: Emerging Business Intelligence' Wiley Publication, ISBN: 978-1-11-814760-3, February 2013

[6] Mark Talabis, Robert McPherson 'Information Security Analytics Finding Security Insights, Patterns, and Anomalies in Big Data' Published by Syngress, ISBN: 978-0-12-800207-0, November 2014

[7] Randy Franklin Smith's 'Cutting through the Hype: What is Big Data Security Analytics' by LogRhythm, webinar ultimate windows security.com

[8] Prof. Yuval Elovici 'Security & Privacy' Information Systems Engineering webinar

[9] Luis Maldonado, Michael Roytman 'Introduction to Big Data Techniques for Cyber Security' New York Information Security meet up January 2015 webinar

[10] Eeyal Kolman RSA 'Machine Learning & Big Data in Cyber Security'08/09/2014, webinar

[11] John Vecchi, Ajay Uggirala 'Revolutionizing Advanced Threat Protection A Modern Three Tired Approach' Solerea, webcast

[12] Greg Masters 'Tapping Big Data Security Analytics to detect breaches, APT & gain actionable Intel' webcast

[13] Carson Kai-Sang Leung, Richard Kyle MacKinnon, Fan Jiang 'Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data' 2014 IEEE International Congress on Big Data, 978-1-4799-50577/14,DOI10.1109/BigData.Congres.2014

[14] Lei Xu, Chunxiao Jiang 'Information Security in Big Data: Privacy and Data Mining' 2169-3536, VOLUME 2, 2014 IEEE