Leveraging Big Data & AI-ML for Security Analytics

NIC-CERT's endeavor towards a predictive cyber security approach

NIC has been a prominent target for cyber-attacks. The sheer volume of government applications, websites, services and databases, hosted and managed by NIC makes it a very lucrative targets for cyber threat actors including nation state actors. NIC has been adopting a layered defense approach for mitigating these attacks, with state-of-the-art technology. As modern technology evolved with much more enhanced attack detection and mitigation capabilities, the attackers also evolved and they started chaining multiple exploits, leading to a multi-vector and multi-stage attack.

In a world with technologies powered by AI-ML, the modern threat and attack landscape have undergone a massive change. To tackle this changing landscape, it is imperative to leverage the very same AI-ML to extract crucial insights and analytics, so as to enhance the overall cyber security posture and be better prepared to detect and respond to attacks, as early as possible.

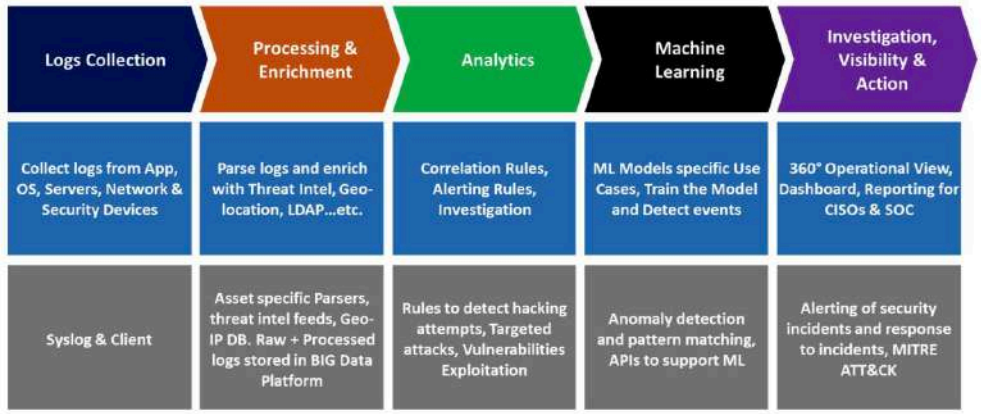
Some of the attacks which involve zero-day vulnerabilities, are even more difficult to detect through conventional security solutions. In order to tackle these changing dynamics in the attacker tactics and techniques, NIC-CERT has embarked on an ambitious initiative for designing and commissioning a robust AI-ML based cyber analytics platform. The platform can be leveraged to spot certain cyber attacks at an early stage and help in improving the security posture of NIC.

R.S. Mani
Dy. Director General
& HoG
rsm@nic.in

Hariharan M
Scientist-C
hariharan.m@nic.in

Gaurav Kansal
Scientist-C
gaurav.kansal@nic.in

▼ Security Analytics Platform - Modules



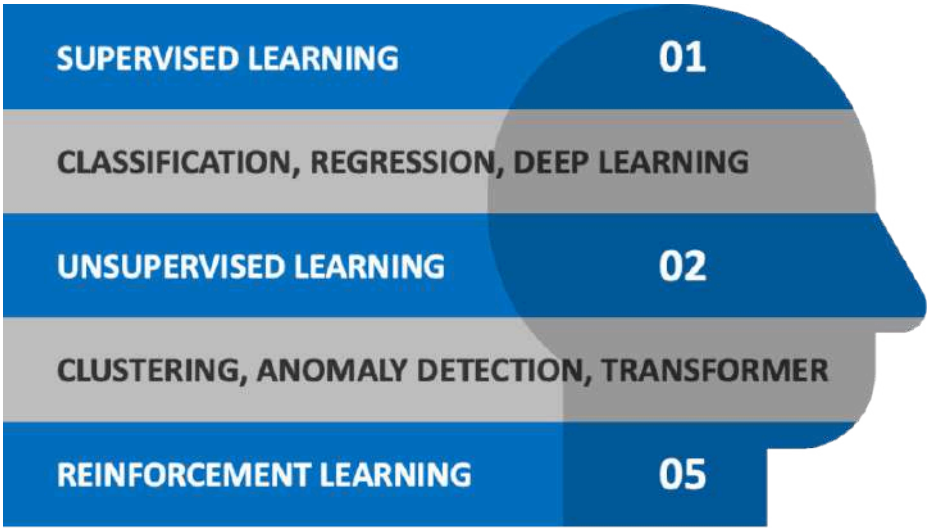
Security Analytics Platform

The security analytics platform is envisioned to handle data in the scale of petabytes and it should be scalable. In this context, Elastic Search and Hadoop can be used as the backend data lake. The elastic search can facilitate the correlation/ alert rules, dashboards and analytics. Whereas, Hadoop can facilitate the machine learning analysis, through additional tools like python, spark. the primary source of data to be ingested into the platform would be the logs generated by various devices, servers, endpoints, applications, websites and services. The logs may be collected from various sources across the Government ICT Infrastructure connected to NICNET and the logs shall be processed and enriched with additional details (like Geo-location, IP/Domain Reputation, etc). The processed logs will then be analyzed on the analytics platform using various correlation and security rules. In addition to this, a machine learning model will also process the logs and will try to identify various anomalies and suspicious patterns in the logs. Multiple Machine Learning models may be integrated into the security analytics Platform, each ML Model will have AI-ML Models for Security Analytics the capability to train and learn, where by it attains certain level of maturity over a period of time. Once the ML Model attains the maturity level, it can spot much more advanced and complex attacks, which may not be spotted by the traditional rule based SIEM platforms.

AI-ML in Advanced Security Analytics and Threat Detection

AI-ML has become the buzzword in recent times. Most of the new technology products claims to leverage AI-ML in one way or the other. Inspite of all the buzz and being touted as the next big thing in the technology evolution, the journey towards achieving successful results through AI-ML is an arduous task; Especially, when it comes to cyber security and threat/attack detection, it would require billions of data events to train the model appropriately, so that it can achieve a certain degree of accuracy.

The classification model under supervised learning can be built around knowledge of known classifier objects such as IP addresses, domain names, network object interactions, and other data points, which are extracted from the logs. This can further be used to build various classification models which can be tested and adopted based on classification accuracies and relevance. Unsupervised learning can be leveraged for better grouping of clusters, where various clustering algorithms need to be used to identify and quantify data relationships from data and meta-data extracted from the logs. Deep Learning neural networks can be used for predicting anomalies in the data set gathered from various log sources. One of the key focus areas of the security analytics platform is to transform the security detection from reactive to proactive i.e., aid in predictive analytics.



Features of the Security Analytics Platform

Some of the key features of the platform are as follows:

- Central Aggregated Log Management Platform
- Web and Security Analytics
- Visibility on Attacker Activity
- Detect/Predict Anomalies or Attacks at an early stage
- Incident Response, Threat Hunting & Threat Intelligence
- Facilitate troubleshooting of website/ application issues
- Dashboard & Reporting

Tactical Insights & Security Posture

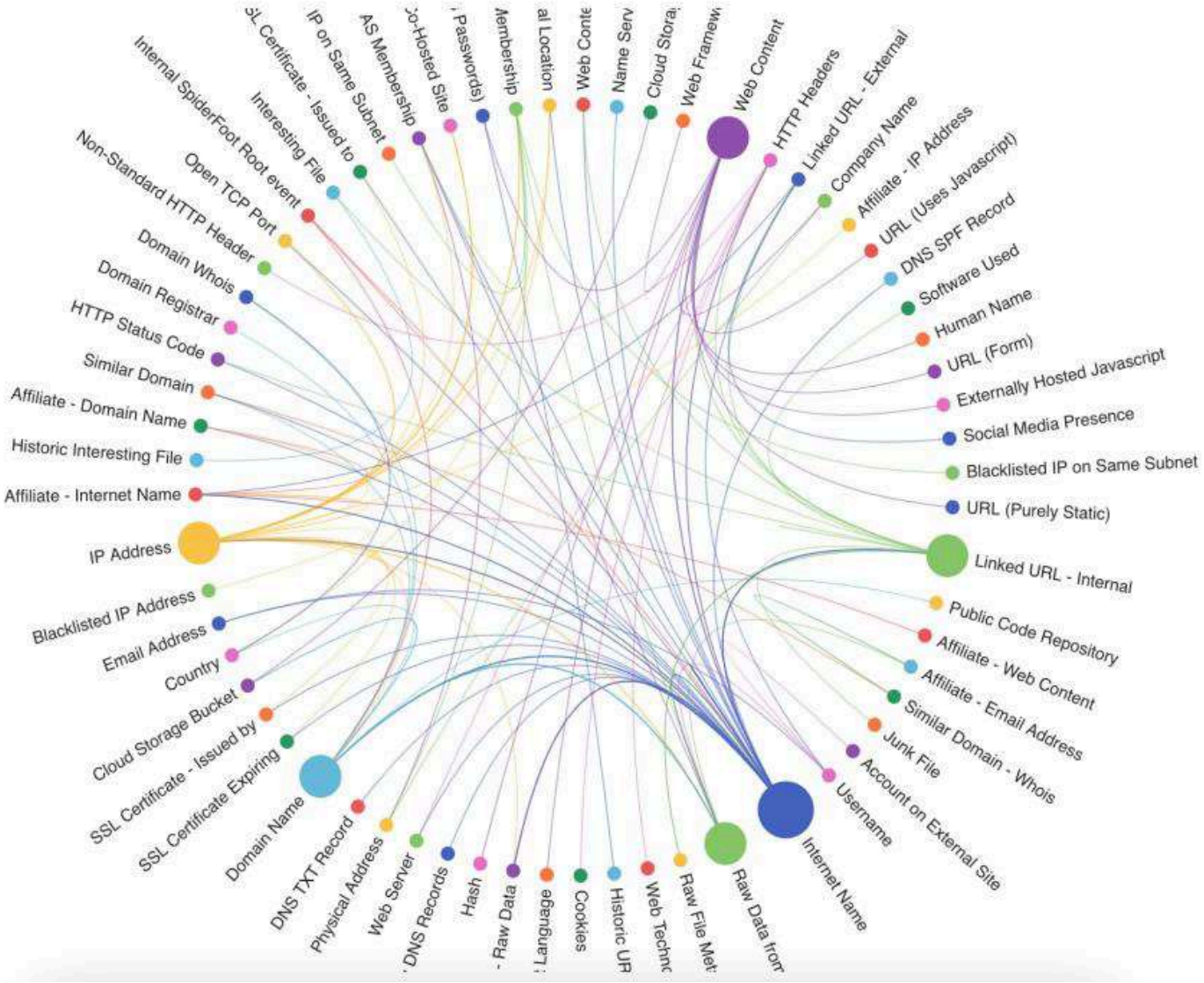
From an ICT perspective, the logs of a system are literally a piece of recorded history of what happened on the system, when it happened, this information can be further inferenced to identify how the specific event happened and why it happened. Considering an organization like NIC, which hosts thousands of websites, applications and not to mention the lakhs of ICT devices spread across the country, collection and aggregation of logs from these devices in itself poses a huge challenge. But if we overcome the challenge and are able to aggregate the data, then the insights that could be derived from the aggregated logs would be invaluable. Since, its practically not possible for a human to physically check and investigate each log event, this is where automated security analytics and machine learning comes into picture; Together, the ML and Security Analytics can quickly sift through billions of log events and filter out those events which could be of interest from a security perspective and it could further extrapolate the relationship between a particular event and a whole plethora of other related data sets, thereby providing tactical insights essential for strengthening our cyber security posture.

The security analytics platform can also provide key web analytics on the site traffic, visitor stats, suspicious hits, etc. The insights generated from one log source can also be correlated with another log source to check for any similarities. For example, an attacker who has attempted to hack into one state government website, was again found to be attempting to hack another central government ministry's website. This is where the analytics platform, will try to inter-relate both the attacks based on various features and attributes, and further the model would try to learn the techniques adopted by the attacker for launching the attack. The learning would then be ingrained into the model and it could train itself for detecting similar such attacks in the future.

Key Benefits of the Security Analytics Platform

The security analytics platform is powered by a massive data lake at the backend, which is essentially a repository of log data collected from various sources. The platform can be leveraged to ask various questions by querying the underlying data to get necessary information. In addition to this, the platform can also offer the following key benefits:

- Huge cost savings in the range of hundreds of crores, which would have been incurred in a corresponding commercial platform
- Security incidents can be identified quickly and action can be taken before any major damage could be done



▲ Extrapolation of Relationships between datasets

- Round the clock visibility can help in identifying which vulnerabilities/loop holes, are being exploited, this information can aid developers/ administrators to fix them quickly
- Can aid in the troubleshooting of issues in Govt. websites/ICT Infrastructure, the platform can also reduce the time to identify and fix the issue
- As logs from multiple NDCs and States are to be ingested into the security analytics platform, the Machine Learning Models, can get exposed to a vast, varied and more unique data events, which can aid in training the models to achieve a much higher level of accuracy
- Ministry/State/Project specific Dashboard and Reporting view, for up-to-date analytics and security posture status

Conclusion

Logs form an important part of an ICT system. All supported ICT systems should be configured to generate and store logs. It is advisable to store the logs in a central logging server, which is independent of the log source. The logs should be configured to capture crucial details like timestamp, source, destination, request, port, protocol, username, etc. The most important aspect is the timestamp, it is essential that all ICT systems within NIC are synchronized with the same time stamp from the central NTP server. If time stamps are not synchronized, then the very purpose of logging may be defeated. Once we collect and aggregate logs from multiple sources, then the AI-ML plays a vital role in fetching key insights from the logs. These insights can further contribute to policy making and other decision support systems. Moreover, it enhances the visibility of what is happening around in the ICT infrastructure and when this visibility combined with the insights, it can become a formidable tool to strengthen the overall security posture of NIC and the government at large.

For further information, please contact:
Hariharan M
Scientist-C
National Informatics Centre
CGO Complex, Lodhi Road
New Delhi - 110 003
Email: hariharan.m@nic.in, Phone: 011-22907465