DATA ANALYTICS FOR OPTIMISING CYBER AND DATA CENTRE OPERATIONS

CHANG Xuquan Stanley, SIM Sze Liang, WONG Ming Qian

ABSTRACT

A wealth of data is generated by the Ministry of Defence's IT networks which can be analysed to improve cyber threat detection and data centre operations.

In cyber defence, detection algorithms have advanced from static rules to machine-learning algorithms that can harness the rich amount of network data available to detect low signature anomalies in the environment. Similarly, statistical analysis on infrastructure logs can derive patterns in system utilisation and user behaviour to gain insights into increasingly complex operating environments, pre-empt incidents and optimise resource allocation.

This article shares how DSTA applies data analytics to enhance efficiency and effectiveness in cyber defence and data centre operations.

Keywords: data analytics, cyber defence, data centre, IT operations analytics

INTRODUCTION

Data analytics is about deriving insights from data. Through mathematical, statistical and machine learning methods, patterns can be discovered to present businesses and other operations with a more relatable view of their data for decision making.

DSTA implements data analytics for the Ministry of Defence (MINDEF) in a number of areas. In the air and maritime domains, analytics is used to augment situational awareness with real-time interpretation of movement patterns and detection of anomalous activity. Analytics is also applied in a variety of Enterprise IT areas such as finance, procurement, logistics and human resource. The rich data sets accumulated over extended periods of operation are analysed to provide support in areas like budget optimisation, fraud detection, supply risk management and staff engagement.

This article describes how DSTA uses data analytics in the applied areas of cyber defence and data centre operations.

DATA ANALYTICS IN CYBER AND DATA CENTRE OPERATIONS

Cyber defence and data centre operations are becoming increasingly challenging to manage due to: (a) increasing criticality of IT; (b) growth in complexity and number of systems; and (c) increasingly sophisticated cyber adversaries and threats.

Traditional incident and event management tools have served adequately in the detection, notification and reporting of events. However, they need to be calibrated manually to detect anomalies and to analyse correlated events and trends.

Data analytics can be implemented to automate the derivation of such insights. Its capabilities include advanced search and indexing techniques that allow effective correlation of events and analysis of statistical patterns. Machine learning methods are also applied to discover topological relationships and establish behavioural baselines.

As such, the existing wealth of data can be mined. The data made available for analytics are: (a) machine data such as utilisation and activity logs from servers and networked devices; (b) network data; and (c) synthetic data, which is data generated by probing the system with simulated test cases.

The application of these data in cyber defence and data centre operations can be categorised into: anomaly detection, discovery of hidden patterns and insights, and optimisation of resources.

Anomaly Detection

Statistical and machine learning methods are used to trend and forecast system utilisation and behavioural patterns continuously, and build baselines that also enable the detection of low-signature anomalies with high fidelity.

By correlating such anomalies across system configurations, compliance checks and security events, preventive measures can be taken to avert the potential onset of performance degradation or halt the progress of cyber attacks. The usage of analytics thus reduces the reliance on tacit knowledge and individual competencies to identify risks, and construct an action plan which would be unsustainable in the face of growing system complexities and manpower constraints.

System utilisation and web access patterns are also correlated to understand, and subsequently anticipate the impact of user activities on the performance of applications. These insights allow data centre operations to prepare for planned user activities effectively or determine the possible causes of an unplanned surge in utilisation.

Anomalies can also be indicative of new exploits or freshly compromised assets in the networks that warrant further investigation by incident response teams. Examples of such anomalous events include reconnaissance activities by external entities who are attempting to gain insights into the network, and computers infected with viruses that are attempting to perform unauthorised actions.

Machine learning algorithms, both supervised and unsupervised, are also being applied to pick up events that exhibit similar behaviours from past anomalous events or deviations. Supervised learning algorithms are used to pick up domain name resolution requests to suspicious domains. Unsupervised learning algorithms such as k-means clustering are used to categorise network and machine data into “normal” and “anomalous” clusters. These machine learning algorithms are commonly used in the cyber domain, especially in the detection of new threats as there is no known information of the threat that can be used to identify it with certainty.

Alternatively, known abnormal behaviours can also be used in analytics to facilitate the categorisation of observed patterns into “normal” and “anomalous” groups based on the patterns’ similarities to known attributes of abnormal behaviours. This approach can be used to supplement anomaly detection in the initial phase of defining the baseline of the environment, where either insufficient time has lapsed to build a reliable baseline or where it is undesirable to assume that learned behaviour is normal by default.

Discovery of Hidden Patterns and Insights

Advanced analytics algorithms are used for indexing, searching and correlating large data sets to discover hidden patterns, relationships and insights that are normally difficult for a human to perceive.

One such example is the analysis of time intervals between Internet requests originating from machines within the networks. Malware often needs to contact its command and control (C2) servers on the Internet to receive further instructions. These requests usually occur in very regular intervals as they are controlled by a programme. This regular pattern is illustrated in Figure 1. By comparison, human-initiated web surfing requests are random with irregular intervals between requests as shown in Figure 2. The standard deviation and entropy between the differences in the time intervals are calculated, and low standard deviations and entropy values give indications that these requests are occurring in a periodic manner. Using this analysis, machines infected with malware that were contacting C2 servers periodically have been detected in the past.

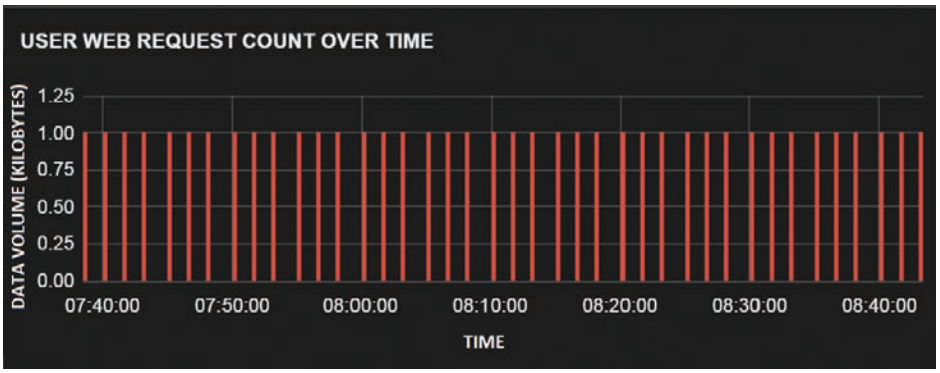

Figure 1. Regular network traffic patterns of malware beaconing

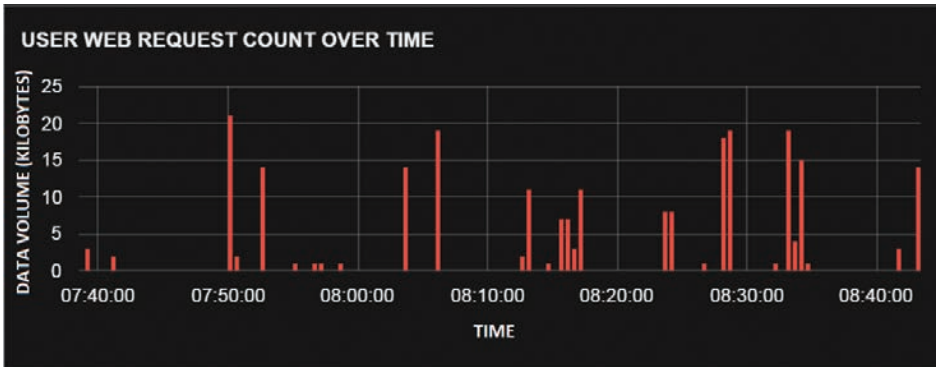

Figure 2. Irregular network traffic pattern of web surfing behaviour

Analytics algorithms are also used to identify cyber attack campaigns by analysing their intrusion indicators, such as the source of the attacks and malware used, against an intrusion attributes database. Hutchins, Cloppert, and Amin (2011) defined a kill chain model[1] to describe the intrusion phases of a cyber attack and identify indicators that link individual attacks to a campaign. Using information derived from the analysis, cyber defence teams can also determine the tactics, techniques and procedures (TTP) of the attackers, allowing them to stay ahead of the attacks.

For data centre operations, advanced searching and indexing capabilities are used to perform root cause analysis for incidents. This is a critical yet time-consuming activity during the resolution of incidents. A common problem is the difficulty in determining the root cause event among a cluster of events that follow after it.

For example, the failure of a common email gateway would eventually result in transactional failures being reported by most applications that serve user workflows. That would generate one alert for each application service for each time it attempts to send an email. Among all of those alerts, only one alert would have been generated for the root cause. Data analysis could be used to establish similarities in time, volume and textual patterns between alerts, thus removing large amounts of noise and making it easier to identify anomalous events that are unique to the time period leading up to the incident.

Analytics is also being used to discover new relationships between events. For example, an application error and a shortage of disk space may be considered as two separate events that are resolved independently. Yet, the shortage of disk space may have been caused by application errors which generate large amounts of logs within a short period of time. Analytics is being applied to discover such correlations and causal relationships, improving both assessment and reaction to events.

Optimisation of Resources

Analytics has become more pertinent in optimising data centre resources in recent years. This can be attributed to the increasing use of virtualisation technologies that enable resource sharing and live migration of workload.

Within the data centre, a requirement placed on server virtualisation and storage platforms is to provide the analytics and automation required to redistribute workload dynamically for optimal system resource usage. This also enables the overall distribution of virtualised application servers to adapt constantly to changes in utilisation patterns and find placement on physical servers that can best serve their needs. For data storage, analytics identifies frequently accessed data continuously for placement on storage resources where it can be served with the best throughput.

Analytics also supports "right-sizing" by utilising historical and projected patterns to anticipate the correct quantities of resource allocation, enable reclamation and redistribution of resources to cope with short-term surges, and moderate changes in demand by optimising resources within existing capacity.

Beyond short-term optimisations, predictive analytics is also used to model the impact that projected requirements may have on existing capacity and anticipate growth requirements.

During this analysis, availability and performance requirements are considered alongside anticipated changes in resource demands as a result of the utilisation and growth patterns of existing deployments. With the inclusion of planned deployments within the capacity planning model, any resulting shortage in capacity and timeframe can guide the precise and timely acquisitions of additional capacity.

CHALLENGES

The following points present challenges that need to be addressed in order to realise the potential of analytics for cyber defence and data centre operations.

Deriving Relevant Insights

Before analytics can be implemented, a problem statement has to be defined to determine the appropriate data and methods to use. It is usually not straightforward to frame a problem statement for analysis. For example, questions like "how to reduce operating costs" or "which servers are infected with malware" are natural questions that arise but are too generic to initiate immediate analysis. Apart from having familiarity with the context of the question, framing a problem statement requires staff to be technically proficient in diverse domains such as IT infrastructure, cybersecurity, statistics and mathematics in order to identify suitable metrics and methods that can derive the required insights. Acquiring such competencies requires an organisational commitment of time and resources.

In practice, where a desired insight is framed at a level that is too obscure to initiate analysis, a thought process to reduce the original question recursively into smaller intrinsic queries helps to make the relevant data and approach more apparent.

Reliability of Analytic Results

The reliability of both descriptive and predictive analysis remains fundamental to the effective use of analytics. In data centre operations where the problem is more defined, numerous analytical tools are available in the market which provide non-traditional means to collect relevant data and analyse them for commonly required insights. This necessitates evaluation of the results produced by these products aside from their technical specifications or capability. However, the challenge lies in validating the accuracy of these results.

In practice, the validation of descriptive analytics results is straightforward as in the case of incident resolution and avoidance, as erroneous analysis is usually obvious and remediable on hindsight. However, in the case of predictive analytics such as resource optimisation or capacity planning, analytics recommendations are used to serve as inputs to the traditional planning process for a period of time before it may be deemed sufficiently reliable to replace the traditional process itself.

CONCLUSION

This article has highlighted some ways which data analytics is used to optimise cyber defence and data centre operations. Data analytics is a significant game-changer that allows more effective application of insights. From the strengthening of security posture to enhancement of user experience, a tangible impact has been made on MINDEF's networks.

Data analytics has also resulted in less downtime due to better predictive capabilities, while improved insights have enabled swifter and more effective actions against cyber threats and service outages.

In addition, data analytics augments the optimisation of resource allocation in data centre operations, enabling rapid application delivery in a more cost–effective approach.

To stay ahead of sophisticated cyber attacks, data analytics has become an important tool to pick up hidden indicators of these threats. As hackers come up with new TTPs, new measures have to be developed and this is made possible by the valuable insights gained from analytics.

Together, a more secure yet effective IT experience can be delivered to MINDEF while incurring less effort and potentially lower operating costs through the use of data analytics.

REFERENCES

Dua, S., & Du, X. (2014). *Data mining and machine learning* in *cybersecurity*. CRC Press: Boca Raton, FL.

Hutchins, E. M., Cloppert, M. J., & Amin, R. M. (2011). Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. In J., Ryan (Ed.), *Leading issues in information warfare & security research* (pp. 78 – 104). Reading, UK: Academic Publishing International Ltd.

Jacobs, J., & Rudis, B. (2014). *Data-driven security: analysis, visualization and dashboards.* John Wiley & Sons: Indianapolis, IN.

Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. *Proceedings of Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen,* 4. GI/ITG-Workshop MMBnet 2007, Hamburg, Germany.

Singhal, A. (2007). *Data warehousing and data mining techniques for cyber security.* Springer: New York.

ENDNOTES

[1] A kill chain is a systematic sequence of events performed by an adversary to select and engage a target to achieve a desired effect. In describing a cyber attack, Hutchin, Cloppert and Amin defined the kill chain as seven phases of activity comprising: (a) reconnaissance to select the target; (b) weaponisation of the attack payload; (c) delivery of the payload; (d) exploitation of the target; (e) installation of trojan or backdoor into the target; (f) establishing command and control channel to control the target; and (g) execution of actions to achieve objectives.

BIOGRAPHY

CHANG Xuquan Stanley is a Manager (Cybersecurity) working on the development and application of analytics for cyber threat detection. Stanley graduated with a Bachelor of Engineering (Computer Engineering) degree with First Class Honours from the National University of Singapore (NUS) in 2006. He further obtained a Master of Science (Defence Technology and Systems) degree from Temasek Defence Systems Institute in 2010 as well as a Master of Science (Computer Science) degree from the Naval Postgraduate School, USA, in 2011.

SIM Sze Liang is a Manager (InfoComm Infrastructure) overseeing data centre virtualisation and private cloud computing initiatives for the Ministry of Defence's (MINDEF) Corporate IT (CIT) infrastructure. Sze Liang has also been actively involved in the modernisation of IT infrastructure and leading implementations for infrastructure consolidation and system automation. In addition, he is exploring the use of analytics to augment incident management and capacity planning for virtualised infrastructure. Sze Liang graduated with a Bachelor of Engineering (Computer Engineering) degree with Honours from NUS in 2009.

WONG Ming Qian is a Senior Engineer (InfoComm Infrastructure) managing and implementing analytics capabilities across MINDEF's CIT infrastructure. He is also involved in the enhancement of existing analytics dashboards to improve functionality and usability of analytical tools and systems. Ming Qian graduated with a Bachelor of Computing (Computing Science) degree from NUS in 2006.