C.W. Axelrod, Int. J. of Design & Nature and Ecodynamics. Vol. 11, No. 3 (2016) 268–274

ANALYZING BIG, MIDSIZE, AND SMALL DATA FOR APPLICATION SECURITY

C.W. AXELROD
Delta Risk LLC, USA

ABSTRACT

Organizations collect huge amounts of security intelligence and yet analysts fail to anticipate many attacks leading to data breaches, denials of service, identity theft, fraudulent use of systems and data, and other nefarious activities. Analysts mostly learn of incidents from third parties, such as law enforcement and payment-card processing companies. Could it be that they do not have available the right level and mix of data? We describe how one might optimize the collection and analysis of security information and event management data, particularly as they apply to securing computer applications. It is argued that this optimization can be achieved by combining big, midsize, and small data and running them through appropriate analytical methods.
Keywords: attacks, big data, cloud computing, exploits, midsize data, preventative measures, security intelligence, small data, surveys, threats, vulnerabilities.

1 1INTRODUCTION

We have seen much discussion about whether cloud computing really differs from regular IT outsourcing and whether cloud service providers can be trusted to keep sensitive data secure. Today, both these areas have hit their respective strides. The rapid growth of the number and business volumes of providers demonstrates a general acceptance of the cloud service model. Concerns about security and privacy are abating as more research, experience, conferences, certifications, and vendor tools for both cloud-computing security and protection of big data become available. On the other hand, we are seeing a staggering growth in the number, size, and significance of data breaches, many of which only become known to victims after notification by third parties [1], suggesting that current internal monitoring and alerting tools are grossly inadequate.

This change in how the demand for, and supply of, computing services and data analysis are being met is not only apparent from growth statistics, but also from a rapid shift toward innovative computer processors and data storage devices and away from devices and software that support traditional data centres [2]. Mainframe computers, dedicated networks, and slow software development cycles have supported computing for half a century. But these resources are rapidly giving way to data centres containing thousands upon thousands of small processors working in unison and running applications created using agile processes such as continuous software integration [3], delivery [4] security [5] and, vulnerability management [6]. At the extreme, we are seeing hundreds to thousands of deployments per day [7]. Data collection, storage, aggregation, analysis, and visualization tools have appeared recently to meet the need for speedy processing of vast amounts of data being generated in the Cloud.

Such rapid evolution has led many information security professionals to embrace big data, and the recent ability to analyze huge data troves using predictive analytical tools and cloud computing

WIT CONFERENCES
This paper is part of the Proceedings of the International Conference on Big Data (Big Data 2016)
www.witconferences.com

ISSN: 1755-7437 (paper format), ISSN: 1755-7445 (online), http://www.witpress.com/journals
DOI: 10.2495/DNE-V11-N3-268-274

services, as the answer to anticipating cyber threats, attacks, and misuse; preventing damage from such attacks; and fixing software vulnerabilities and weaknesses. However, the increasing numbers of more sophisticated and successful cyber attacks casts doubt upon whether we are in fact meeting cyber security challenges or falling further behind.

This is not to deny that big data and new analytical tools hold great promise in producing new and valuable information about threats, exploits, attacks, and vulnerabilities. They do. But big data alone are not enough to solve the many problems that still confront analysts. Consequently, there is a need to supplement big data with so-called 'small data' obtained from surveys and human-to-human interactions [8], and with what the author calls 'midsize data', which are typically obtained from logs created by security products, such as firewalls and intrusion detection and prevention systems (IDPS), as well as from instrumentation programmed into applications, as described in [9].

2 DATA SIZE AND CHARACTERISTICS

The categorization of data into big, midsize, and small appears to be quite simple, as in Table 1, which shows the sizes and typical data for big, midsize, and small data. However, there are significant overlaps across categories particularly between midsize and big data. Some large companies generate so much information that it can be considered big data. In a 2013 version [10], it was claimed that the Hewlett Packard enterprise was generating one trillion events per day. In [9], it is suggested that big data tools might be used to analyze the data thrown off by the sensors built into applications. In general, we can think of big data emanating from the Cloud, midsize data being generated internally within an organization, and small data deriving from internal surveys and discussions. However, there are exceptions to this classification such as online surveys generating big data.

In Fig. 1, we see the three major sources of security intelligence data and indicate whether they are collected and analyzed, reported, and responded to in real time or batch mode. In today's environment, the need for real-time responses is high due to the speed with which cyber attacks take place. Fortunately, many modern analytical methods enable real-time processing and decision-making

Table 1: Data by size and type.

Data	Size	Types of information produced
Big	Terabytes to petabytes	Threat advisories Exploit advisories Incident (successful attack) reports Unusual activity reports Forensics reports
Midsize	Gigabytes to terabytes	Reports and alerts of attempted attacks Reports and alerts of successful attacks Unusually activity alerts and reports Forensics reports
Small	Kilobytes to gigabytes	Context Business value Uncertainty and risk postures Different perspectives on criticality of systems and data

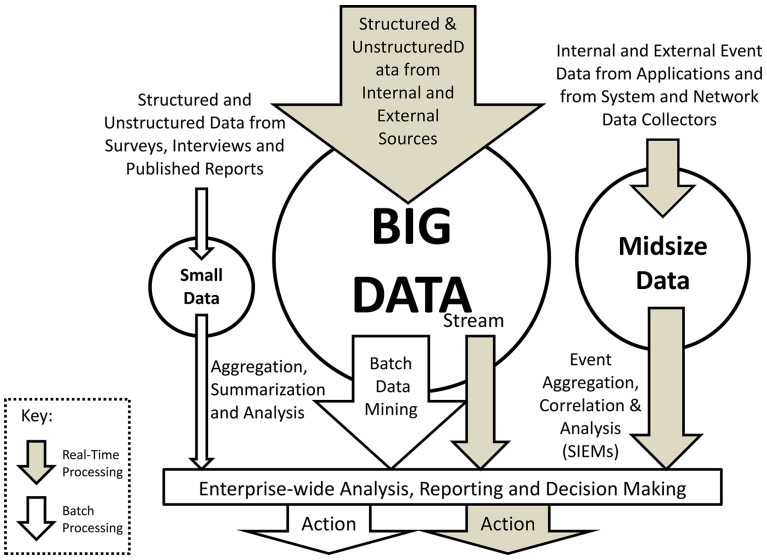

Figure 1: Real-time and batch data collection, processing, and reporting. (Source: Axelrod [11], by permission)

with big and midsize data. Small data results are used to create context, which usually changes only occasionally, so that use of batch mode is more acceptable.

3 TRADITIONAL FOCUS ON TRAFFIC DATA

While the software security movement has grown substantially over the past two decades—particularly with the rapid growth of OWASP (Open Web Application Security Project) and the U.S. Department of Homeland Security support of their 'Build Security In' website—we have not seen universal acceptance of the importance of application security within the cyber security domain. In this author's view, this is because relatively few information security professionals have a software development background. Instead, because many cyber security professionals come out of network and system engineering and operations, much of the focus of cyber security has been on traffic—messages and transactions over the Web or in the Cloud.

While there is no doubt that sifting through huge volumes of traffic data seeking patterns of malware and other nefarious activities does yield helpful results and is augmented significantly with big-data storage and analysis, it is clear from the increasing number of successful attacks that this approach is wanting. Furthermore, with legislators, regulators, and much of the public insisting on the need for encryption to protect private and sensitive data, encryption and anonymity are seen by many as the answer to many security issues, despite claims that relatively few actual data compromises have taken place over networks and that most successful attacks are against data 'at rest'.

It is because traffic data are so voluminous and so valuable for deriving customer purchasing proclivities and the like, that so much attention is being given to cyber security threats and attack patterns extracted from big data using those same big-data analysis techniques [12]. A review of tools available and their applicability can be found in [13]. Furthermore, threat data are becoming more readily available from such sources as Soltra Edge [14], which is a relatively new service offering actionable real-time threat information.

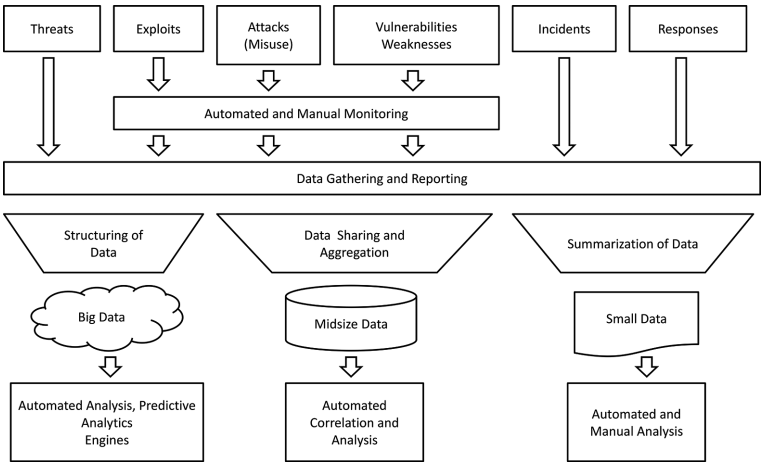

Figure 2: Data processing methods as threats morph into incidents.

As we proceed from threats to exploits to incidents and to responses, we find that we are moving from big data through midsize data to small data, as illustrated in Fig. 2. For the most part, generalized security tools, such as firewalls, antivirus software, and the like, are being used even as we become more specific with respect to the particular context in which applications operate. We now look at some new initiatives, sponsored by OWASP, which may help to resolve these issues.

4 PURSUING APPLICATION-GENERATED DATA

In the past, there have been normative discussions, such as in [15], advocating the instrumentation of application software to obtain more specific application-generated data, as well as suggestions for functional security testing [16] to ensure that systems do not behave as they shouldn't. Now, we are beginning to see guidance as to how to implement these approaches. In particular, the AppSensor project by OWASP, documented in [9], has produced guidelines as to what to implement and is providing specific methods in order to achieve these goals. Incorporation of AppSensor capabilities provides attack detection and real-time defensive response to attack-aware software.

AppSensor's defining characteristics are as follows:

- Detection of a selection of suspicious and malicious events
- Use of the results of the detection to identify attacks centrally
- Selection of a predefined response to attacks
- Execution of these responses

In Table 2, we see what AppSensor purports to do and what it is not meant to do. It should be noted that AppSensor only operates effectively when software has been developed using a secure software development lifecycle and is current with respect to updates and patches so that known weaknesses and vulnerabilities will have already been addressed.

Some of the key benefits of the AppSensor approach, versus other application-protection approaches, are as follows [12]:

- Knowledge about business logic
- Knowledge about roles and permissions of users

Table 2: AppSensor capabilities.

AppSensor Does ...	AppSensor Does NOT ...
... detects users seeking vulnerabilities	... analyzes application source code
... helps prevent attackers from finding known weaknesses	... stops exploits of which the organization is already aware
... adds attack detection and prevention capabilities to applications	... examines application in run-time environment
... adds detection and prevention capabilities from other systems	... provides benefits if a secure development lifecycle is not in place
...works with detailed business-specific data related to a particular application or set of applications	... performs dynamic patching
... determines an attack where a user is stepping through a multi-step business process in the wrong order	... acts as a 'silver bullet'
... understands when there are changes in user role or in classification of the data	
... understands when users exceed user-specific action thresholds	

- Ability to arrive at informed decisions about misuse
- Very low false-positive rates compared to other approaches
- Ability to identify and stop attackers.

This information can only become available if there are attack-detection sensors already built into the applications.

With this information at hand, appropriate responses can be formulated and invoked. Such responses need to be determined in advance and related to different forms of attack, including those that have not been seen previously.

5 BLURRING OF LINES AMONG DATA TYPES AND ANALYTICAL TOOLS

It is becoming evident that there is increasing spill-over among data types as well as among various types of tools used for analyzing different formats (e.g. structured vs. unstructured), types (e.g. threats, exploits, vulnerabilities), and sizes (i.e. big, midsize, small) of data.

In particular, we are seeing ever larger amounts of data being generated within organizations so that these data are moving into the realm of big data. But even if the data generated internally do not match the sheer quantities of data being farmed in the Cloud, there are real advantages in using the data storage and analytical tools within organizations. In addition, as internal data become too large to store and process internally, it increasingly makes more sense to use cloud services to handle such data.

These improvements in efficiency and effectiveness are good news for cyber security, which has been greatly stymied in its progress due to the need for massive resources to collect, store, and process security data and derive meaningful metrics. In the *AppSensor Guide* [9, p. ix], we learn that:

> 'OWASP AppSensor captures so much data that ... it becomes possible to apply big data analytics to security ... This is an area for AppSensor's future development ...'

However, later in the report [9, p. 4], we read that:

> 'Application-specific attack detection does not need to identify all invalid usage, to be able to determine an attack. There is no need for "infinite data" or "big data" in this approach'.

While there seems to be some contradiction here, the *AppSensor Guide* appears to be saying that one should be selective in collecting attack data, but eventually the amount of data may become sufficiently large to apply techniques used today for analyzing big data to AppSensor-generated data. This supports the idea that the boundary between midsize and big data is blurred rather than sharp.

However, despite some equivocation, we are seeing throughout the literature trends toward more cyber security data being monitored and collected and the appearance of impressive big-data storage and analysis tools. Ultimately, we need to determine the value of the security intelligence and whether the results are worth the effort and cost. In [17], Axelrod suggests that more valuable cyber security data often are more difficult and costly to obtain and analyze, but frequently worth the investment. Until recently, desired storage, processing, and analytical capabilities were not readily available at reasonable cost. With the rapid evolution of big-data technologies and projects such as AppSensor, we might finally be seeing the much-needed progress to produce valuable and timely security information.

6 CONCLUSIONS

Many researchers seem to be so enthralled by the prospects of big data and big-data analytics that they appear to believe that big data hold the answers to all their unanswered questions. However, as has been pointed out by some big-data experts, big data alone cannot provide all the answers and needs to be supplemented with small data to provide the much needed context, as was described in [8].

In this paper we have argued that, when it comes to cyber security data and metrics, there is a raft of midsize data that supplies information specific to an organization's systems and networks and the software that it runs. In order to get a complete picture of an organization's cyber security ecosystem, researchers need to combine and analyze all three of these data categories.

When it comes to analyzing the data, there are specific tools that apply directly to each data category (i.e. big, midsize, and small). However, as the boundaries between data categories blur, it is becoming economically and technically feasible, as well as desirable, to apply modern analytical tools more broadly across data categories.

We stand on the brink of a veritable revolution in the collection and analysis of, and response to, data from cyber security events. It is hoped that such innovative approaches will overcome the asymmetry between attackers and defenders so that defenders can effectively overcome many of the advantages that attackers have heretofore held.

7 FURTHER RESEARCH

While it has not been discussed specifically in this paper, one should consider the feasibility of expanding sharing of cyber security information, not only at the general threat level (as with Soltra Edge [14], for example) but also at the organizational level. It would be interesting to see how AppSensor attack-detection and defensive-response data might be shared across organizations within an industry, sector, country, or region. Sharing attack data should allow organizations to better anticipate potential attacks against them, using predictive analytics. The aggregation and analysis of response data will result in organizations being better informed as to how best they might deal with them. Coincidentally, there is an emerging big-data analytics area called 'prescriptive analytics' that is intended to facilitate this type of analysis. It is designed to let you know '... the best way to get to where you want to be' [18].

This would be a very interesting, as well as challenging, area for further investigation as it combines addressing the technical hurdles of data sharing with national and international economic and political considerations.

REFERENCES

[1] Verizon Enterprise, *2015 Data Breach Investigations Report*, Verizon, 2015, available at http://www.verizonenterprise.com/DBIR/2015/

[2] Metz, C., Dell. EMC. HP. Cisco. These tech giants are the walking dead, *Wired Magazine*, October 2015, available at http://www.wired.com/2015/10/meet-walking-dead-hp-cisco-dell-emc-ibm-oracle/

[3] Duvall, P.M., Matyas, S. & Glover, A., *Continuous Integration: Improving Software Quality and Reducing Risk*, Addison-Wesley, 2007.

[4] Humble, J. & Farley, D., *Continuous Delivery: Reliable Software Releases through Build, Test and Deployment Automation*, Addison-Wesley, 2010.

[5] Hoff, J. & Chapple, M., *Securing the SDLC for Dummies*, John Wiley, 2014.

[6] Kandeck, W., *Vulnerability Management for Dummies,* 2nd edn., John Wiley, 2015.

[7] Kim, G., Behr, K. & Spafford, G., *DevOps Guide—Selected Resources to Start Your Journey*, IT Revolution Press, 2015, available at http://www.delphix.com/wp-content/uploads/2015/09/delphix-ar-itrev-devops-guide.pdf

[8] Peysakhovich, A. & Stevens-Davidowitz, S., How not to drown in numbers, *Sunday Review, The New York Times*, May 2, 2015.

[9] Watson, C., Groves, D. & Melton, J., *AppSensor Guide: Application-Specific Real Time Attack Detection & Response, Version 2.0*, OWASP (Open Web Application Security Project), July 2015.

[10] Murthy, P., Bharadwaj, A., Subrahmanyam, P.A., Roy, A. & Rajan, S., *Big Data Working Group: Big Data Taxonomy*, Cloud Security Alliance, September 2014.

[11] Axelrod, C.W., Actionable security intelligence from big, midsize and small data. *ISACA Journal*, **1**, pp. 44–50, 2016.

[12] Watson, C., Coates, M., Melton, J. & Groves, D., Creating attack-aware software applications with real-time defences. *CrossTalk*, **24(5)**, pp. 14–18, 2011.

[13] Watson, C., Chan, J., Hall, M. & Ven der stock, A., *OWASP Automated Threat Handbook—Web Applications, Version 1.01*, OWASP, October 2015.

[14] DTCC (Depository Trust Clearing Corporation), *Soltra Edge, the First Industry-Driven Threat Intelligence Sharing Platform Now Generally Available, Easy-to-Use and Free to License*, DTCC Press Release, December 3, 2014.

[15] Axelrod, C.W., Creating data from applications for detecting stealth attacks, *CrossTalk*, **24**(**2**), pp. 17–21, March/April 2011.

[16] Axelrod, C.W., The need for functional security testing, *CrossTalk*, **24**(**5**), pp. 19–24, September/October 2011.

[17] Axelrod, C.W., Accounting for value and uncertainty in security metrics. *ISACA Journal*, **6**, 2008.

[18] Pratt, M.K., Five things you need to know: prescriptive analytics, *CIO Magazine*, p. 20, December 2014/January 2015.