WHITE PAPER

Big Data for Advanced Threat Protection

Key Criteria for Cutting Through the Clamor

Sponsor: Trend Micro

Author: Mark Bouchard

AimPoint Group

keeping IT on target

Introduction

One of today's hottest topics is the potential for big data technologies and techniques (Big Data) to help address information security (InfoSec) problems. At a macro level, this makes perfect sense. Big Data is proven in other areas of IT/business (e.g., web site analytics and fraud management), there's no shortage of security related data to work with, and there are plenty of solution areas to pursue – not the least of which is advanced threat protection to help combat targeted attacks and the continued rise of malware.

From a practical perspective, however, Big Data for InfoSec has some challenges. The biggest obstacle is that cost and complexity of a Big Data implementation put this approach out of reach for all but the largest enterprises with the most well-funded and well-staffed IT departments. Far from a show-stopper, this impediment just means that until the segment matures sufficiently – that is, to the point that infrastructure costs come down, best practices solidify, and appropriately skilled personnel become more available – most enterprises will obtain Big Data for InfoSec capabilities indirectly. Rather than investing in Big Data management infrastructure and associated analytic tools and applying these resources to high-profile security problems themselves, they will instead benefit from similar investments made by security solution providers intent on enhancing their core offerings.

The next challenge, then, is cutting through the clamor – all of the noise and inflated claims of solution providers latching onto the Big Data for InfoSec bandwagon. Navigating the hype to establish a solution's true capabilities relative to Big Data requires close attention to detail. Just because a security provider has a lot of data at its disposal and can successfully manage it doesn't mean their solution is necessarily capable of producing actionable intelligence, or, for that matter, of doing so at the speeds necessary to effectively mitigate advanced threats.

This paper further examines the Big Data for InfoSec scenario and enumerates essential criteria for enterprises to use when evaluating advanced threat protection and other security solutions purporting to take advantage of Big Data to achieve superior effectiveness.

Big Data Makes Big Sense for Security

Big Data involves using specialized technologies and techniques to collect, coordinate, store, and analyze truly massive amounts of related and perhaps even disparate data to uncover insights and patterns that would otherwise remain obscured. The use of "Big" reflects that these solutions and approaches were born out of the needs of the likes of Amazon, Google, and Yahoo!, and, unlike ordinary data mining products, are intended to support Internet-scale operations.

Leveraging Big Data for information security purposes not only makes sense but is necessary. To begin with, Big Data implementations have proven successful in other areas of IT/business, and there is no reason to expect this success won't convey to InfoSec use cases. This is particularly true given the plethora of data available for analysis. 50 gigabytes is a relatively conservative estimate for the security log and event data collected daily by organizations with more than 1,000 employees. More than a few IT departments collect a terabyte or more – and that doesn't even account for operations and other types of non-security data that is nonetheless useful for the additional context that it provides. For security solution providers, the volume of data is even more staggering. Trend Micro, for example, processes over 6 terabytes daily with its Smart Protection Network.

Leveraging Big Data for information security purposes not only makes sense but is necessary.

Most important, however, are the significant, pervasive, and intractable security problems Big Data can help address.

For instance, it is well documented that malware and targeted attacks are on the rise, and that most organizations are struggling to effectively combat these threats.[1] Conventional countermeasures continue to do their jobs, often admirably, but are simply not up to the challenge of server-side polymorphism, customized attacks, and other advanced tactics. Big Data can help in this case by generating substantially better models of normal/good behavior and revealing deviations from these patterns to more thoroughly and accurately identify malicious activity.

Another issue Big Data can help resolve is the shockingly high frequency of incidents where readily available data shows clear signs of an attack being underway, yet it still goes undetected. For example, according to Verizon's 2012 Data Breach Investigations Report, post-mortem analysis of the incidents under review revealed that 84% of victim organizations had evidence of a breach in their log files, but that this data either went unnoticed or was not acted upon. This situation is indicative of the fact that organizations are routinely overwhelmed by the sheer volume of security data being collected and/or are not analyzing it sufficiently. Big Data's value proposition in this case is the promise of far more efficient data management capabilities and embedded analytics.

Additional Big Data for InfoSec use cases include:

- Enabling better/adaptive management of access privileges via role analytics
- Enabling increased monitoring to compensate for the loss of direct control accompanying such trends as consumerization, bring-your-own-device (BYOD), and the adoption of cloud services
- Delivering true, risk-based intelligence by providing a better understanding of the business value of individual components and resources

Indirect Exposure to Big Data for InfoSec Will Dominate (For Now)

Reasons organizations may be reluctant to make substantial investments in Big Data for InfoSec over the near term include the following:

- IT departments already suffering from a "deluge of data" may be disinclined to pursue strategies that require managing (and making sense of) even more data.
- Depending on the problems an organization is looking to address, the data sources at their disposal may not be sufficient (i.e., "big" enough).
- "Big" data management is not a trivial pursuit. At the same time that the underlying technology may be unfamiliar (e.g., Hadoop, NoSQL, analytic databases, BASE vs. ACID, etc.), critical requirements pertaining to performance, scale, and diversity of data types must be met and balanced. Architecture choices must be made for individual technologies (e.g., does a columnar database make sense?) as well as the solution overall (e.g., to what extent are data and processing services centralized, or should they remain distributed?).
- "Big" data analysis is not a trivial pursuit. Many of the previous issues/considerations apply to this domain too. In addition, there is the challenge of finding personnel with the appropriate skills (i.e., a fusion of InfoSec and data science), as well as general lack of mature, packaged analytics targeted at security-specific problems.
- Cost-benefit metrics are non-existent, leaving internal champions with an uphill battle when it comes to justifying related investments.

The situation for security solution providers, however, is completely different. They already have many of the requisite resources: terabytes of useful data from global technology footprints spanning thousands of enterprises; additional data derived from threat and vulnerability research departments; skilled personnel from the same; and, experience managing mountains of data. For some providers, in fact, Big Data for InfoSec is not a new endeavor (e.g., Cisco, Trend Micro, and Websense with their respective global intelligence networks/services). In addition, investments can be spread over an entire customer base. Moreover, making these investments is a no-brainer; it's what's needed to maintain a competitive advantage.

The net result is that, at least for the next few years, direct investments in Big Data for InfoSec by enterprises will be limited primarily to large organizations with well-funded and staffed IT departments. Exposure for everyone else will be indirect – as a back-end capability that feeds into the other security solutions they already have or acquire.

More specifically, most instances will involve threat protection infrastructure and/or security management solutions enhanced by a global intelligence network leveraging Big Data. For this latter group, a dedicated, enterprise-specific implementation will involve either steady evolution of an existing platform (e.g., SIEM), or a concentrated investment once Big Data for InfoSec sufficiently matures and commoditizes.

Beware Of The Big Data Bandwagon

Regardless of when and to what extent enterprises eventually make direct investments in Big Data for InfoSec, security solution providers have clearly recognized the value of doing so now. One predictable result is that there's a lot of "noise" in the market. And although numerous solution providers are making claims of using Big Data to enhance their offerings, the unfortunate reality is that there is considerable variation in terms of the degree to which these claims are true. For example, simply having a NoSQL or Hadoop based data management

infrastructure is not the same as also: (a) running a pile of powerful analytics against it to obtain unique and valuable insights, and (b) delivering actionable intelligence to customers in a timely manner.

Accordingly, the remaining sections of this paper enumerate essential evaluation criteria enterprises can use to help cut through the noise and ensure selection of an effective solution – at least with regard to its Big Data implementation. The context for these criteria is a Big Data enabled global intelligence service used to enhance on-premise threat protection and/or security management infrastructure. However, the majority of what's covered should be applicable to other implementations and use cases too.

Data, Data … and More Data

A big data solution is nothing without data, and lots of it. Even more important is the type and quality of data. In this regard, it's important to recognize that the threat landscape is global, pervasive, and highly dynamic. Threats are coming from practically everywhere and acting against practically everything, every minute of every day. New, unique variants are emerging at the astounding rate of more than 1 per second, and are using a diverse set of channels to spread (e.g., web sites, email, files, mobile apps, removable media, social media, etc.).[2] Having a chance of finding them – not to mention doing so consistently – requires casting a broad net, one that provides coverage (i.e., collects data) across many of the following domains and dimensions of IT:

- Networks – actual network activity is an essential complement to system logs
- Device types – fixed and mobile endpoints, servers, common networking devices, and high-profile security platforms are just a starting point
- Platforms – MAC, Linux, iOS, and Android are just a few of the operating systems to consider besides Windows
- Applications – data from mobile, social networking, and legacy apps should not be overlooked
- Enterprises and geographic regions – single enterprise, US-centric data pales in comparison to cross-enterprise data gathered from organizations worldwide
- Functional areas within IT/Security – identity, vulnerability, event, threat, and operations management systems are especially rich data sources

The bottom line is that although "intelligence" derived from a global footprint of CPE-based threat protection products (e.g., antivirus, ids/ips, etc.) is good, to the extent that a solution has deeper visibility within individual enterprises, that is even better. Only with a combination of cross-organization and organization-specific intelligence will enterprises be able to effectively thwart not just broad spectrum threats but targeted attacks as well.

Data Management Infrastructure

Data management infrastructure – or the components needed to collect, aggregate, store, and process Internet-scale data sets – can be a tricky topic. Far too many solution providers hold up their investment in Hadoop, MapReduce, NoSQL, and similar technologies as definitive evidence that they in fact have a Big Data for InfoSec solution. But data management infrastructure is only one piece of the bigger puzzle – and one, it can be argued, that ultimately ranks behind breadth and depth of data sources, analytics, and skilled personnel in terms of overall importance.

Only with a combination of cross-organization and organization-specific intelligence will enterprises be able to effectively thwart not just broad spectrum threats but targeted attacks as well.

Another point to consider is that although this particular set of technologies has an early and impressive following in the Big Data arena, they are certainly not the only options, and quite possibly not even the best ones. For example, limits to the parallel scalability of Hadoop are increasingly becoming obvious. Also, let's not forget the value of substantial enterprise investments in and familiarity with ordinary relational technologies.

Rather than succumbing to the glimmer of the Big Data technology du jour, IT/Security managers should instead press potential solution providers for details such as the following when it comes to the underlying data management infrastructure that is involved:

- How does it blend both relational and non-relational technologies to maximum effect?
- How are dynamic scalability and high performance achieved?
- How is diversity of data types accounted for?
- What architectural decisions/features have been implemented to help optimize for search, cor relation, and other forms of analysis?
- In what ways has it evolved (i.e., been improved) over the past couple of years?
- What techniques are employed to help moderate the amount of data being stored?

Data Analysis Capabilities

Scanning the web site or data sheet content for the typical Big Data for InfoSec solution will reveal that "advanced technologies including Big Data analysis techniques" or "data mining algorithms" are employed to get the drop on advanced threats and generate actionable insights. Invariably, however, precious few additional details are provided. To some extent this "black box" approach is defensible, at least for publicly available resources.

After all, for a legitimate solution, the "further details" represent significant intellectual property and the special sauce that makes it different/better than other solutions in the market.

However, such obscure treatment of this all-important topic can also hide the fact that a solution is seriously deficient in its actual capabilities – perhaps that it's focused primarily on the movement and management of data, and not so much on enabling in-depth analysis to uncover anything other than surface-level insights. IT/Security managers, therefore, at least in their private enterprise-to-provider communications, should press for further details in this area. Specific capabilities and characteristics to pursue include the following:

- What core analysis techniques are employed (e.g., search, visualization, data mining, machine learning, etc.)?
- How much emphasis is placed on machine learning (which focuses on reproducing known knowledge) versus data mining (which focuses on discovery of previously unknown properties of data), and why?
- How much emphasis is placed on developing a better understanding of how threats behave, versus creating better models of known good behavior and detecting deviations from them, and why?
- What sorts of unique and customized tools are employed to set the solution apart?
- What tools and techniques are used to account for unstructured data?

- What steps have been taken to ensure employed techniques yield insights that are not only actionable but available in time to do some good (versus after the fact)?
- In what ways has innovation been pursued to help stay ahead of the bad guys?

The bottom line is that the management of data is not the same as data analysis, and a solution lacking adequate data analysis capabilities will be next-to-worthless in the fight against modern malware and targeted attacks.

Skilled Personnel

Whether it's a direct or indirect implementation of Big Data for InfoSec, appropriately skilled personnel are a critical element for success. One of the challenges in this regard is the relative shortage of such staff. Specific skill sets that are needed include:

- Data management expertise, for architecting and operating high performance, highly scalable data collection and processing infrastructure
- Data analysis expertise, for statistical analysis, data mining (etc.), and developing new analytics
- Threat analysis expertise, for identifying, modeling, and staying on top of "bad behaviors" and other techniques employed by attackers

Each of these is fairly rare to begin with, and finding personnel with all three sets of capabilities is nearly impossible. A second challenge, then, is getting a team of such experts to work together smoothly, particularly when each wants to be in control and their individual agendas are not always compatible. For example, data management experts tend to emphasize scalability over everything else, while accuracy is paramount for the data scientists and speed of detection often rules for the threat analysts. This suggests the need for a fourth member of the staff: the ringleader – a proficient manager with an above-average understanding of each domain of expertise who can balance competing objectives to maintain an effective solution.

When evaluating Big Data for InfoSec solution providers, IT/Security managers should not be reluctant to thoroughly investigate the quantity, quality, and relative mix of personnel associated with the offering. The inability or unwillingness for a solution provider to satisfactorily respond definitely constitutes a red flag.

Process Matters Too

Having already discussed people and technology, that brings us to the third leg of the triumvirate necessary for any good InfoSec solution: process. Think of this as another acid test for prospective solution providers. If there is no evidence of well-defined processes underpinning their offering, then they're apparently relying on chance to generate useful threat intelligence – and there's no point paying for that.

The bottom line is that the management of data is not the same as data analysis, and a solution lacking adequate data analysis capabilities will be next-to-worthless in the fight against modern malware and targeted attacks.

Specific processes to look for and/or evaluate include, at a minimum, the following:

- Data collection and aggregation – including accounting for secondary sources
- Data elimination – not all data needs to be (or should be) stored indefinitely
- Data analysis – what approach do analysts use to methodically process available data
- Threat collaboration – what steps do analysts take to identify relationships among threats within an attack
- Threat validation – to help eliminate false positives
- Threat response – see below

Threat Response

The value of new threat intelligence is directly tied to how it's put to use. At a minimum, there needs to be a notification capability that proactively pushes details of newly discovered threats, ideally along with best practice recommendations and tactical guidance on how to counteract them, to key personnel within customer organizations. For those enterprises willing to embrace such an approach, an even stronger defense can be established based on the solution provider pursuing a combination of manual and automatic generation of corresponding signatures, rules and configuration settings and dynamically delivering them to customer-operated protection technologies (e.g., AV, FW, IPS, etc.), across all platforms (e.g., physical, virtual, cloud, and mobile). Of course, speed is again an important factor to evaluate (and balance against accuracy). Knowledge and defense against new threats is significantly less valuable if it arrives after the threat itself!

Automation

A substantial percentage of organizations that have implemented SIEM solutions suffer from what they describe as a "data deluge." There's simply too much data being collected and not enough time to do anything meaningful with it. Another way to look at this situation, is that these solutions lack sufficient automation (beyond the data collection process) to get tasks like analysis and response activities done efficiently – and, for that matter, in a timeframe that preserves their value. To be clear, this involves capturing and embedding potentially complex processes, knowledge, and learning capabilities. All of this is extremely difficult, which is the reason most SIEM platforms continue to be deficient in this regard.

This same challenge also applies, arguably even more so, when using Big Data for global threat intelligence solutions. Consequently, evaluators should press solution providers for details in this area. Which pieces of the overall solution are automated, and for those that aren't, why? To what extent is automation a point of emphasis for the solution provider? What are they doing to capture/create new analytics to help automate the data analysis process? The goal is to gain sufficient insight to be assured the solution provider can truly handle a mountain of data.

Conclusion

By themselves, traditional countermeasures are not faring well against today's advanced threats. One approach that holds significant promise, though, is the use of threat intelligence networks that take advantage of Big Data techniques and technologies to enhance existing threat protection solutions – to basically give them up-to-the-minute insight into new threats so they can effectively defend against them.

The challenge for enterprises considering this or any other type of Big Data for InfoSec implementation is cutting through the clamor to separate the truly effective solutions from those that are merely capitalizing on the Big Data phenomenon. In this regard, evaluators are encouraged to focus on the selection criteria outlined herein, paying particular attention to: breadth and depth of data sources, data analysis capabilities, and the presence and commitment of appropriately skilled personnel.

Footnotes:

(1) See any of these representative resources: *Trends in Targeted Attacks*, Trend Micro, 2011; *Symantec Internet Security Threat Report* (Vol 17, April 2012), or *Websense 2012 Threat Report*

(2) Research from Trend Micro indicates a release rate of greater than 1 new threat per second. The Symantec Internet Security Threat Report (Vol. 17, April 2012) indicates that 403 million unique malware variants were discovered in 2011, suggesting a release rate of more than 11 threats/second.