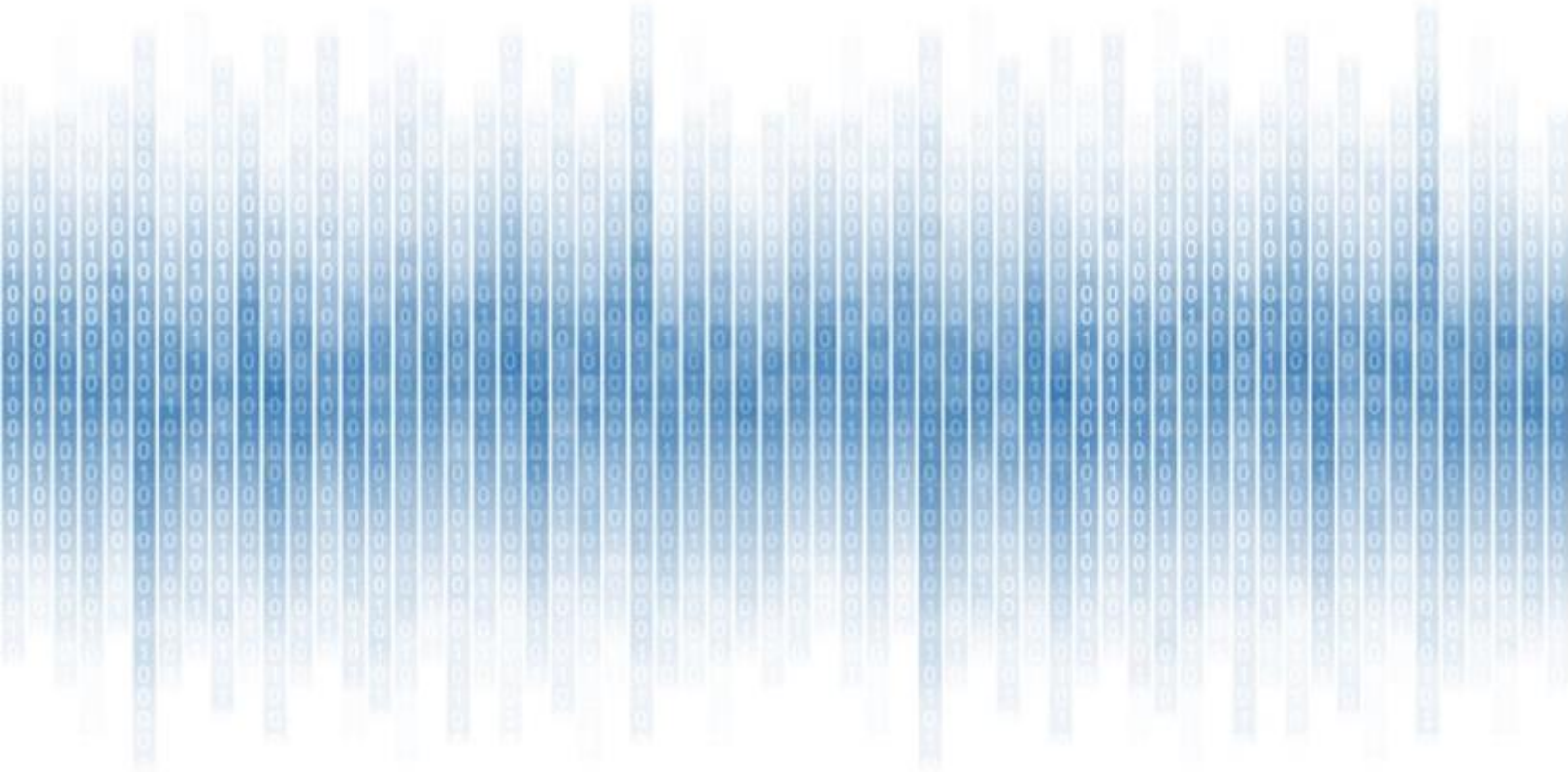
cloud security alliance℠



Big Data Working Group

Big Data Analytics for Security Intelligence

September 2013

Contents

Acknowledgments 4
1.0 Introduction 5
2.0 Big Data Analytics 6
2.1 Data Privacy and Governance 10
3.0 Big Data Analytics for Security 10
4.0 Examples 11
4.1 Network Security 11
4.2 Enterprise Events Analytics 12
4.3 Netflow Monitoring to Identify Botnets 13
4.4 Advanced Persistent Threats Detection 14
4.4.1 Beehive: Behavior Profiling for APT Detection 15
4.4.2 Using Large-Scale Distributed Computing to Unveil APTs 16
5.0 The WINE Platform for Experimenting with Big Data Analytics in Security 17
5.1 Data Sharing and Provenance 17
5.2 WINE Analysis Example: Determining the Duration of Zero-Day Attacks 18
6.0 Conclusions 19
Bibliography 21

Acknowledgments

Editors:
Alvaro A. Cárdenas, University of Texas at Dallas
Pratyusa K. Manadhata, HP Labs
Sree Rajan, Fujitsu Laboratories of America

Contributors:
Alvaro A. Cárdenas, University of Texas at Dallas
Tudor Dumitras, University of Maryland, College Park
Thomas Engel, University of Luxembourg
Jérôme François, University of Luxembourg
Paul Giura, AT&T
Ari Juels, RSA Laboratories
Pratyusa K. Manadhata, HP Labs
Alina Oprea, RSA Laboratories
Cathryn Ploehn, University of Texas at Dallas
Radu State, University of Luxembourg
Grace St. Clair, University of Texas at Dallas
Wei Wang, AT&T
Ting-Fang Yen, RSA Laboratories

CSA Staff:
Alexander Ginsburg, Copyeditor
Luciano JR Santos, Global Research Director
Kendall Scoboria, Graphic Designer
Evan Scoboria, Webmaster
John Yeoh, Research Analyst

1.0 Introduction

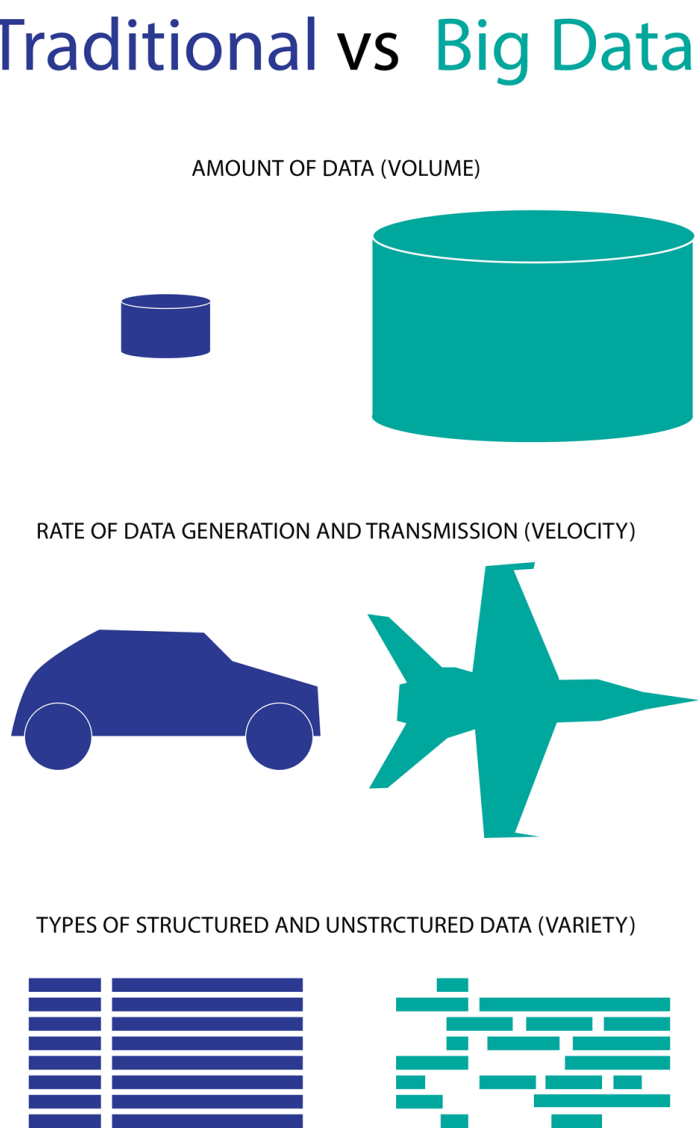


Figure 1. Big Data differentiators

The term *Big Data* refers to large-scale information management and analysis technologies that exceed the capability of traditional data processing technologies.[1] Big Data is differentiated from traditional technologies in three ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety) (Laney, 2001) (Figure 1).

[1] http://gartner.com/it-glossary/big-data/

Human beings now create 2.5 quintillion bytes of data per day. The rate of data creation has increased so much that 90% of the data in the world today has been created in the last two years alone.[2] This acceleration in the production of information has created a need for new technologies to analyze massive data sets. The urgency for collaborative research on Big Data topics is underscored by the U.S. federal government's recent $200 million funding initiative to support Big Data research.[3]

This document describes how the incorporation of Big Data is changing security analytics by providing new tools and opportunities for leveraging large quantities of structured and unstructured data. The remainder of this document is organized as follows: Section 2 highlights the differences between traditional analytics and Big Data analytics, and briefly discusses tools used in Big Data analytics. Section 3 reviews the impact of Big Data analytics on security and Section 4 provides examples of Big Data usage in security contexts. Section 5 describes a platform for experimentation on anti-virus telemetry data. Finally, Section 6 proposes a series of open questions about the role of Big Data in security analytics.

2.0 Big Data Analytics

Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics (Figure 2).

[2] http://www-01.ibm.com/software/data/bigdata/

[3] http://www.nytimes.com/2012/03/29/technology/new-us-research-will-aim-at-flood-of-digital-data.html

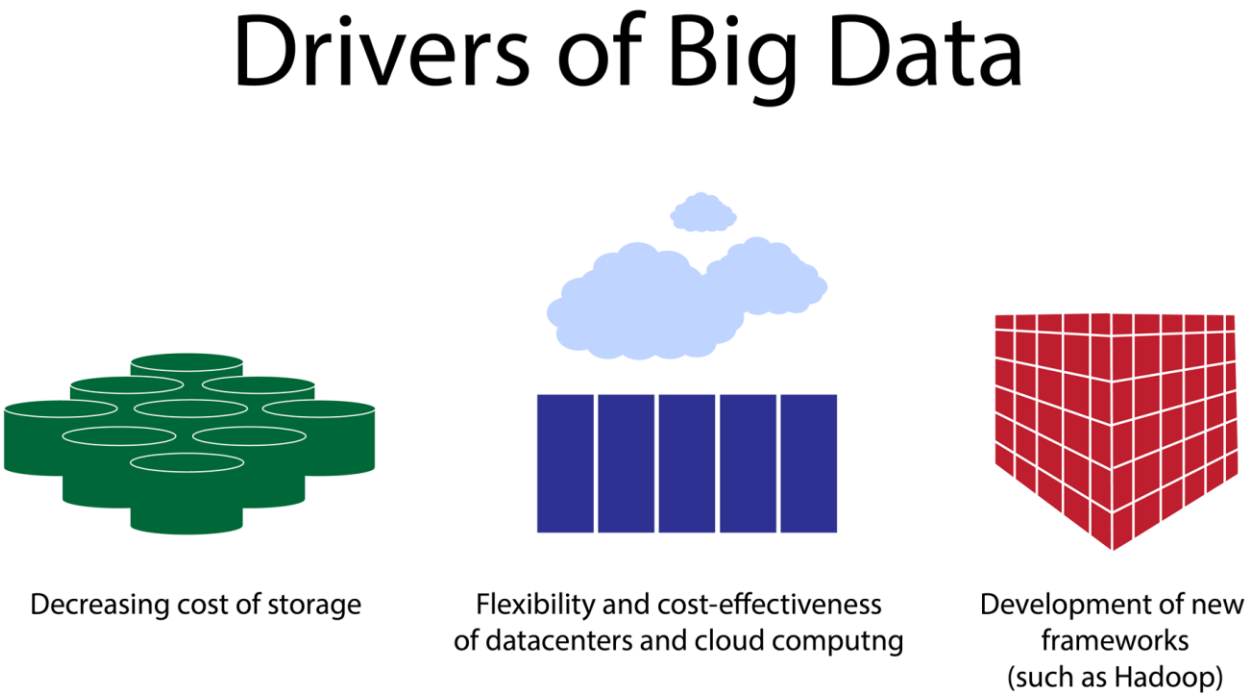


Figure 2. Technical factors driving Big Data adoption

1. Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.
2. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics.
3. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

Big Data technologies can be divided into two groups: *batch processing*, which are analytics on data at rest, and *stream processing*, which are analytics on data in motion (Figure 3). Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis; however, Figure 1 still represents the general trend of these technologies.

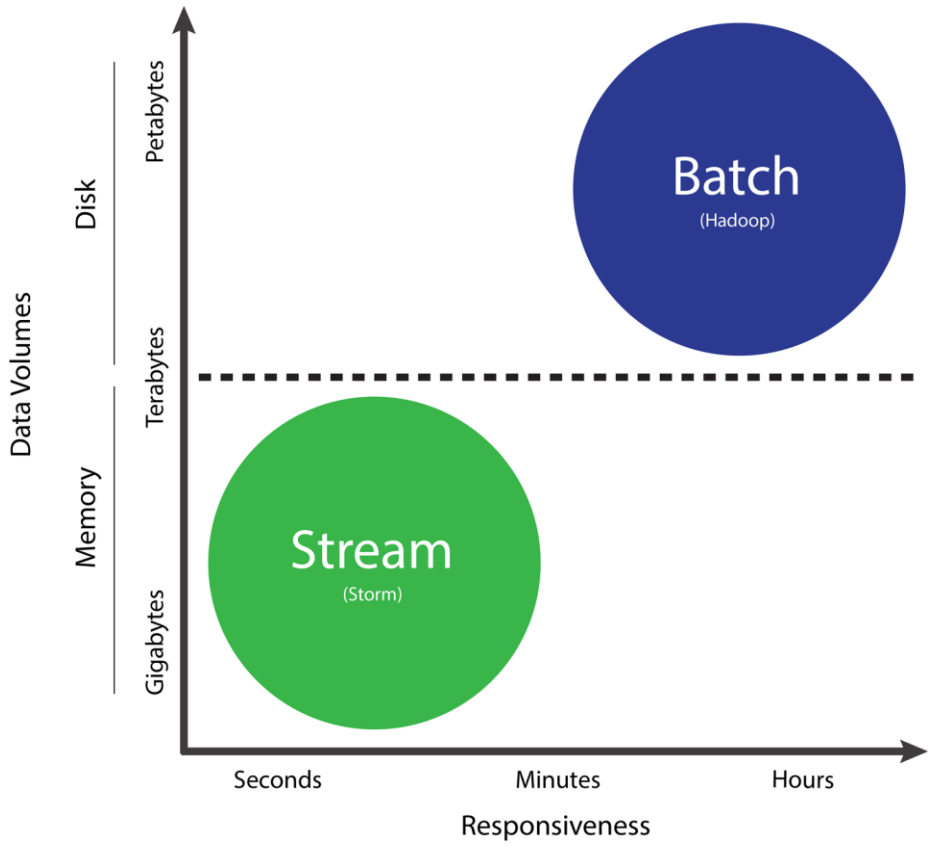


Figure 3. Batch and stream processing

Hadoop is one of the most popular technologies for batch processing. The Hadoop framework provides developers with the Hadoop Distributed File System for storing large files and the MapReduce programming model (Figure 4), which is tailored for frequently occurring large-scale data processing problems that can be distributed and parallelized.

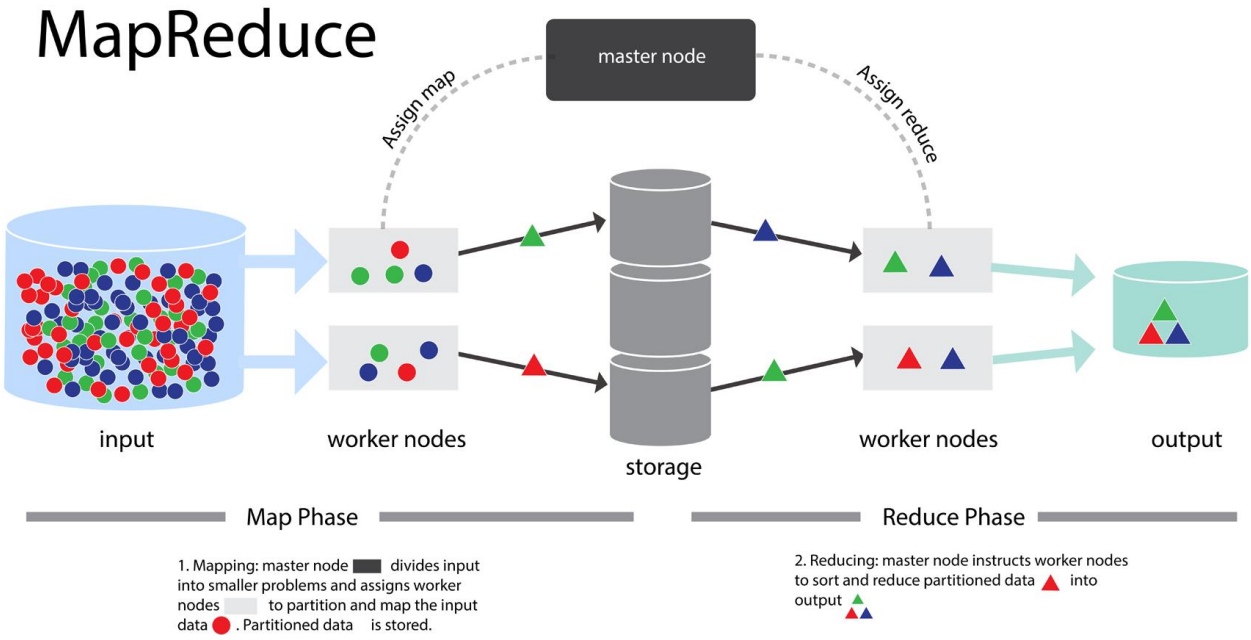


Figure 4. Illustration of MapReduce

Several tools can help analysts create complex queries and run machine learning algorithms on top of Hadoop. These tools include Pig (a platform and a scripting language for complex queries), Hive (an SQL-friendly query language), and Mahout and RHadoop (data mining and machine learning algorithms for Hadoop). New frameworks such as Spark[4] were designed to improve the efficiency of data mining and machine learning algorithms that repeatedly reuse a working set of data, thus improving the efficiency of advanced data analytics algorithms.

There are also several databases designed specifically for efficient storage and query of Big Data, including Cassandra, CouchDB, Greenplum Database, HBase, MongoDB, and Vertica.

Stream processing does not have a single dominant technology like Hadoop, but is a growing area of research and development (Cugola & Margara 2012). One of the models for stream processing is Complex Event Processing (Luckham 2002), which considers information flow as notifications of events (patterns) that need to be aggregated and combined to produce high-level events. Other particular implementations of stream technologies include InfoSphere Streams[5], Jubatus[6], and Storm[7].

[4] http://www.spark-project.org
[5] http://www-01.ibm.com/software/data/infosphere/streams/
[6] http://jubat.us/en/overview.html
[7] http://storm-project.net/

2.1 Data Privacy and Governance

The preservation of privacy largely relies on technological limitations on the ability to extract, analyze, and correlate potentially sensitive data sets. However, advances in Big Data analytics provide tools to extract and utilize this data, making violations of privacy easier. As a result, along with developing Big Data tools, it is necessary to create safeguards to prevent abuse (Bryant, Katz, & Lazowska, 2008).

In addition to privacy, data used for analytics may include regulated information or intellectual property. System architects must ensure that the data is protected and used only according to regulations.

The scope of this document is on how Big Data can improve information security best practices. CSA is committed to also identifying the best practices in Big Data privacy and increasing awareness of the threat to private information. CSA has specific working groups on Big Data privacy and Data Governance, and we will be producing white papers in these areas with a more detailed analysis of privacy issues.

3.0 Big Data Analytics for Security

This section explains how Big Data is changing the analytics landscape. In particular, Big Data analytics can be leveraged to improve information security and situational awareness. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Off-the-shelf Big Data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance, and other fields.

In the context of data analytics for intrusion detection, the following evolution is anticipated:

- 1st generation: Intrusion detection systems – Security architects realized the need for layered security (e.g., reactive security and breach response) because a system with 100% protective security is impossible.
- 2nd generation: Security information and event management (SIEM) – Managing alerts from different intrusion detection sensors and rules was a big challenge in enterprise settings. SIEM systems aggregate and filter alarms from many sources and present actionable information to security analysts.
- 3rd generation: Big Data analytics in security (2nd generation SIEM) – Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating, consolidating, and contextualizing diverse security event information, and also for correlating long-term historical data for forensic purposes.

Analyzing logs, network packets, and system events for forensics and intrusion detection has traditionally been a significant problem; however, traditional technologies fail to provide the tools to support long-term, large-scale analytics for several reasons:

1. Storing and retaining a large quantity of data was not economically feasible. As a result, most event logs and other recorded computer activity were deleted after a fixed retention period (e.g., 60 days).
2. Performing analytics and complex queries on large, structured data sets was inefficient because traditional tools did not leverage Big Data technologies.
3. Traditional tools were not designed to analyze and manage unstructured data. As a result, traditional tools had rigid, defined schemas. Big Data tools (e.g., Piglatin scripts and regular expressions) can query data in flexible formats.
4. Big Data systems use cluster computing infrastructures. As a result, the systems are more reliable and available, and provide guarantees that queries on the systems are processed to completion.

New Big Data technologies, such as databases related to the Hadoop ecosystem and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies will transform security analytics by: (a) collecting data at a massive scale from many internal enterprise sources and external sources such as vulnerability databases; (b) performing deeper analytics on the data; (c) providing a consolidated view of security-related information; and (d) achieving real-time analysis of streaming data. It is important to note that Big Data tools still require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

4.0 Examples

This section describes examples of Big Data analytics used for security purposes.

4.1 Network Security

In a recently published case study, Zions Bancorporation[8] announced that it is using Hadoop clusters and business intelligence tools to parse more data more quickly than with traditional SIEM tools. In their experience, the quantity of data and the frequency analysis of events are too much for traditional SIEMs to handle alone. In their traditional systems, searching among a month's load of data could take between 20 minutes and an hour. In their new Hadoop system running queries with Hive, they get the same results in about one minute.[9]

[8] http://www.darkreading.com/monitoring/a-case-study-in-security-big-data-analys/232602339

[9] http://www.darkreading.com/security-monitoring/167901086/security/news/232602339/a-case-study-in-security-big-data-analysis.html

The security data warehouse driving this implementation not only enables users to mine meaningful security information from sources such as firewalls and security devices, but also from website traffic, business processes and other day-to-day transactions.[10] This incorporation of unstructured data and multiple disparate data sets into a single analytical framework is one of the main promises of Big Data.

4.2 Enterprise Events Analytics

Enterprises routinely collect terabytes of security relevant data (e.g., network events, software application events, and people action events) for several reasons, including the need for regulatory compliance and post-hoc forensic analysis. Unfortunately, this volume of data quickly becomes overwhelming. Enterprises can barely store the data, much less do anything useful with it. For example, it is estimated that an enterprise as large as HP currently (in 2013) generates 1 trillion events per day, or roughly 12 million events per second. These numbers will grow as enterprises enable event logging in more sources, hire more employees, deploy more devices, and run more software. Existing analytical techniques do not work well at this scale and typically produce so many false positives that their efficacy is undermined. The problem becomes worse as enterprises move to cloud architectures and collect much more data. As a result, the more data that is collected, the less actionable information is derived from the data.

The goal of a recent research effort at HP Labs is to move toward a scenario where more data leads to better analytics and more actionable information (Manadhata, Horne, & Rao, forthcoming). To do so, algorithms and systems must be designed and implemented in order to identify actionable security information from large enterprise data sets and drive false positive rates down to manageable levels. In this scenario, the more data that is collected, the more value can be derived from the data. However, many challenges must be overcome to realize the true potential of Big Data analysis. Among these challenges are the legal, privacy, and technical issues regarding scalable data collection, transport, storage, analysis, and visualization.

Despite the challenges, the group at HP Labs has successfully addressed several Big Data analytics for security challenges, some of which are highlighted in this section. First, a large-scale graph inference approach was introduced to identify malware-infected hosts in an enterprise network and the malicious domains accessed by the enterprise's hosts. Specifically, a host-domain access graph was constructed from large enterprise event data sets by adding edges between every host in the enterprise and the domains visited by the host. The graph was then seeded with minimal ground truth information from a black list and a white list, and belief propagation was used to estimate the likelihood that a host or domain is malicious. Experiments on a 2 billion HTTP request data set collected at a large enterprise, a 1 billion DNS request data set collected at an ISP, and a 35 billion network intrusion detection system alert data set collected from over 900 enterprises worldwide showed that high true positive rates and low false positive rates can be achieved with minimal ground truth information (that is, having limited data labeled as normal events or attack events used to train anomaly detectors).

[10] http://www.businesswire.com/news/home/20110802005827/en/Zettaset%E2%80%99s-Security-Data-Warehouse-Enables-Big-Data

Second, terabytes of DNS events consisting of billions of DNS requests and responses collected at an ISP were analyzed. The goal was to use the rich source of DNS information to identify botnets, malicious domains, and other malicious activities in a network. Specifically, features that are indicative of maliciousness were identified. For example, malicious fast-flux domains tend to last for a short time, whereas good domains such as *hp.com* last much longer and resolve to many geographically-distributed IPs. A varied set of features were computed, including ones derived from domain names, time stamps, and DNS response time-to-live values. Then, classification techniques (e.g., decision trees and support vector machines) were used to identify infected hosts and malicious domains. The analysis has already identified many malicious activities from the ISP data set.

4.3 Netflow Monitoring to Identify Botnets

This section summarizes the BotCloud research project (Fraçois, J. et al. 2011, November), which leverages the MapReduce paradigm for analyzing enormous quantities of Netflow data to identify infected hosts participating in a botnet (François, 2011, November). The rationale for using MapReduce for this project stemmed from the large amount of Netflow data collected for data analysis. 720 million Netflow records (77GB) were collected in only 23 hours. Processing this data with traditional tools is challenging. However, Big Data solutions like MapReduce greatly enhance analytics by enabling an easy-to-deploy distributed computing paradigm.

BotCloud relies on BotTrack, which examines host relationships using a combination of PageRank and clustering algorithms to track the command-and-control (C&C) channels in the botnet (François et al., 2011, May). Botnet detection is divided into the following steps: dependency graph creation, PageRank algorithm, and DBScan clustering.

The dependency graph was constructed from Netflow records by representing each host (IP address) as a node. There is an edge from node A to B if, and only if, there is at least one Netflow record having A as the source address and B as the destination address. PageRank will discover patterns in this graph (assuming that P2P communications between bots have similar characteristics since they are involved in same type of activities) and the clustering phase will then group together hosts having the same pattern. Since PageRank is the most resource-consuming part, it is the only one implemented in MapReduce.

BotCloud used a small Hadoop cluster of 12 commodity nodes (11 slaves + 1 master): 6 Intel Core 2 Duo 2.13GHz nodes with 4 GB of memory and 6 Intel Pentium 4 3GHz nodes with 2GB of memory. The dataset contained about 16 million hosts and 720 million Netflow records. This leads to a dependency graph of 57 million edges.

The number of edges in the graph is the main parameter affecting the computational complexity. Since scores are propagated through the edges, the number of intermediate MapReduce key-value pairs is dependent on the number of links. Figure 5 shows the time to complete an iteration with different edges and cluster sizes.

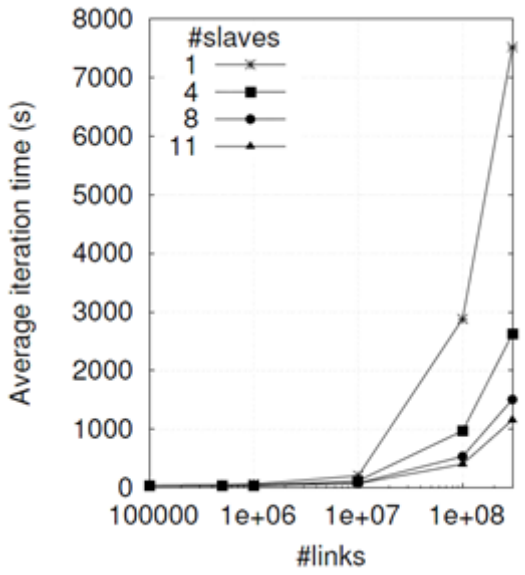


Figure 5. Average execution time for a single PageRank iteration.

The results demonstrate that the time for analyzing the complete dataset (57 million edges) was reduced by a factor of seven by this small Hadoop cluster. Full results (including the accuracy of the algorithm for identifying botnets) are described in François et al. (2011, May).

4.4 Advanced Persistent Threats Detection

An Advanced Persistent Threat (APT) is a targeted attack against a high-value asset or a physical system. In contrast to mass-spreading malware, such as worms, viruses, and Trojans, APT attackers operate in "low-and-slow" mode. "Low mode" maintains a low profile in the networks and "slow mode" allows for long execution time. APT attackers often leverage stolen user credentials or zero-day exploits to avoid triggering alerts. As such, this type of attack can take place over an extended period of time while the victim organization remains oblivious to the intrusion. The 2010 Verizon data breach investigation report concludes that in 86% of the cases, evidence about the data breach was recorded in the organization logs, but the detection mechanisms failed to raise security alarms (Verizon, 2010).

APTs are among the most serious information security threats that organizations face today. A common goal of an APT is to steal intellectual property (IP) from the targeted organization, to gain access to sensitive customer data, or to access strategic business information that could be used for financial gain, blackmail, embarrassment, data poisoning, illegal insider trading or disrupting an organization's business. APTs are operated by highly-skilled, well-funded and motivated attackers targeting sensitive information from specific organizations and operating over periods of months or years. APTs have become very sophisticated and diverse in the methods and technologies used, particularly in the ability to use organizations' own employees to penetrate the IT systems by using social engineering methods. They often trick users into opening spear-phishing messages that are customized for each victim (e.g., emails, SMS, and PUSH messages) and then downloading and installing specially crafted malware that may contain zero-day exploits (Verizon, 2010; Curry et al., 2011; and Alperovitch, 2011).

Today, detection relies heavily on the expertise of human analysts to create custom signatures and perform manual investigation. This process is labor-intensive, difficult to generalize, and not scalable. Existing anomaly detection proposals commonly focus on obvious outliers (e.g., volume-based), but are ill-suited for stealthy APT attacks and suffer from high false positive rates.

Big Data analysis is a suitable approach for APT detection. A challenge in detecting APTs is the massive amount of data to sift through in search of anomalies. The data comes from an ever-increasing number of diverse information sources that have to be audited. This massive volume of data makes the detection task look like searching for a needle in a haystack (Giura & Wang, 2012). Due to the volume of data, traditional network perimeter defense systems can become ineffective in detecting targeted attacks and they are not scalable to the increasing size of organizational networks. As a result, a new approach is required. Many enterprises collect data about users' and hosts' activities within the organization's network, as logged by firewalls, web proxies, domain controllers, intrusion detection systems, and VPN servers. While this data is typically used for compliance and forensic investigation, it also contains a wealth of information about user behavior that holds promise for detecting stealthy attacks.

4.4.1 Beehive: Behavior Profiling for APT Detection

At RSA Labs, the observation about APTs is that, however subtle the attack might be, the attacker's behavior (in attempting to steal sensitive information or subvert system operations) should cause the compromised user's actions to deviate from their usual pattern. Moreover, since APT attacks consist of multiple stages (e.g., exploitation, command-and-control, lateral movement, and objectives), each action by the attacker provides an opportunity to detect behavioral deviations from the norm. Correlating these seemingly independent events can reveal evidence of the intrusion, exposing stealthy attacks that could not be identified with previous methods.

These detectors of behavioral deviations are referred to as "anomaly sensors," with each sensor examining one aspect of the host's or user's activities within an enterprise's network. For instance, a sensor may keep track of the external sites a host contacts in order to identify unusual connections (potential command-and-control channels), profile the set of machines each user logs into to find anomalous access patterns (potential "pivoting" behavior in the lateral movement stage), study users' regular working hours to flag suspicious activities in the middle of the night, or track the flow of data between internal hosts to find unusual "sinks" where large amounts of data are gathered (potential staging servers before data exfiltration).

While the triggering of one sensor indicates the presence of a singular unusual activity, the triggering of multiple sensors suggests more suspicious behavior. The human analyst is given the flexibility of combining multiple sensors according to known attack patterns (e.g., command-and-control communications followed by lateral movement) to look for abnormal events that may warrant investigation or to generate behavioral reports of a given user's activities across time.

The prototype APT detection system at RSA Lab is named *Beehive*. The name refers to the multiple weak components (the "sensors") that work together to achieve a goal (APT detection), just as bees with differentiated

roles cooperate to maintain a hive. Preliminary results showed that *Beehive* is able to process a day's worth of data (around a billion log messages) in an hour and identified policy violations and malware infections that would otherwise have gone unnoticed (Yen et al., 2013).

In addition to detecting APTs, behavior profiling also supports other applications, including IT management (e.g., identifying critical services and unauthorized IT infrastructure within the organization by examining usage patterns), and behavior-based authentication (e.g., authenticating users based on their interaction with other users and hosts, the applications they typically access, or their regular working hours). Thus, *Beehive* provides insights into an organization's environment for security and beyond.

4.4.2 Using Large-Scale Distributed Computing to Unveil APTs

Although an APT itself is not a large-scale exploit, the detection method should use large-scale methods and close-to-target monitoring algorithms in order to be effective and to cover all possible attack paths. In this regard, a successful APT detection methodology should model the APT as an attack pyramid, as introduced by Giura & Wang (2012). An attack pyramid should have the possible attack goal (e.g., sensitive data, high rank employees, and data servers) at the top and lateral planes representing the environments where the events associated with an attack can be recorded (e.g., user plane, network plane, application plane, or physical plane). The detection framework proposed by Giura & Wang groups all of the events recorded in an organization that could potentially be relevant for security using flexible correlation rules that can be redefined as the attack evolves. The framework implements the detection rules (e.g., signature based, anomaly based, or policy based) using various algorithms to detect possible malicious activities within each context and across contexts using a MapReduce paradigm.

There is no doubt that the data used as evidence of attacks is growing in volume, velocity, and variety, and is increasingly difficult to detect. In the case of APTs, there is no known bad item that IDS could pick up or that could be found in traditional information retrieval systems or databases. By using a MapReduce implementation, an APT detection system has the possibility to more efficiently handle highly unstructured data with arbitrary formats that are captured by many types of sensors (e.g., Syslog, IDS, Firewall, NetFlow, and DNS) over long periods of time. Moreover, the massive parallel processing mechanism of MapReduce could use much more sophisticated detection algorithms than the traditional SQL-based data systems that are designed for transactional workloads with highly structured data. Additionally, with MapReduce, users have the power and flexibility to incorporate any detection algorithms into the Map and Reduce functions. The functions can be tailored to work with specific data and make the distributed computing details transparent to the users. Finally, exploring the use of large-scale distributed systems has the potential to help to analyze more data at once, to cover more attack paths and possible targets, and to reveal unknown threats in a context closer to the target, as is the case in APTs.

5.0 The WINE Platform for Experimenting with Big Data Analytics in Security

The Worldwide Intelligence Network Environment (WINE) provides a platform for conducting data analysis at scale, using field data collected at Symantec (e.g., anti-virus telemetry and file downloads), and promotes rigorous experimental methods (Dumitras & Shoue, 2011). WINE loads, samples, and aggregates data feeds originating from millions of hosts around the world and keeps them up-to-date. This allows researchers to conduct open-ended, reproducible experiments in order to, for example, validate new ideas on real-world data, conduct empirical studies, or compare the performance of different algorithms against reference data sets archived in WINE. WINE is currently used by Symantec's engineers and by academic researchers.

5.1 Data Sharing and Provenance

Experimental research in cyber security is rarely reproducible because today's data sets are not widely available to the research community and are often insufficient for answering many open questions. Due to scientific, ethical, and legal barriers to publicly disseminating security data, the data sets used for validating cyber security research are often mentioned in a single publication and then forgotten. The "data wishlist" (Camp, 2009) published by the security research community in 2009 emphasizes the need to obtain data for research purposes on an ongoing basis.

WINE provides one possible model for addressing these challenges. The WINE platform continuously samples and aggregates multiple petabyte-sized data sets, collected around the world by Symantec from customers who agree to share this data. Through the use of parallel processing techniques, the platform also enables open-ended experiments at scale. In order to protect the sensitive information included in the data sets, WINE can only be accessed on-site at Symantec Research Labs. To conduct a WINE experiment, academic researchers are first required to submit a proposal describing the goals of the experiment and the data needed. When using the WINE platform, researchers have access to the raw data relevant to their experiment. All of the experiments carried out on WINE can be attributed to the researchers who conducted them and the raw data cannot be accessed anonymously or copied outside of Symantec's network.

WINE provides access to a large collection of malware samples and to the contextual information needed to understand how malware spreads and conceals its presence, how malware gains access to different systems, what actions malware performs once it is in control, and how malware is ultimately defeated. The malware samples are collected around the world and are used to update Symantec's anti-virus signatures. Researchers can analyze these samples in an isolated "red lab," which does not have inbound/outbound network connectivity in order to prevent viruses and worms from escaping this isolated environment.

A number of additional telemetry data sets, received from hosts running Symantec's products, are stored in a separate parallel database. Researchers can analyze this data using SQL queries or by writing MapReduce tasks.

These data sets include anti-virus telemetry and intrusion-protection telemetry, which record occurrences of known host-based threats and network-based threats, respectively. The binary reputation data set provides information on unknown binaries that are downloaded by users who participate in Download Insight, Symantec's reputation-based security program. The history of binary reputation submissions can reveal when a particular threat has first appeared and how long it existed before it was detected. Similarly, the binary stability data set is collected from the users who participate in the Performance Insight program, which reports the health and stability of applications before users download them. This telemetry data set reports application and system crashes, as well as system lifecycle events (e.g., software installations and uninstallations). Telemetry submission is an optional feature of Symantec products and users can opt out at any time.

These data sets are collected at high rates and the combined data volume exceeds 1 petabyte. To keep the data sets up-to-date and to make them easier to analyze, WINE stores a representative sample from each telemetry source. The samples included in WINE contain either all of the events recorded on a host or no data from that host at all, allowing researchers to search for correlations among events from different data sets.

This operational model also allows Symantec to record metadata establishing the provenance of experimental results (Dumitras & Efstathopoulos, 2012), which ensures the reproducibility of past experiments conducted on WINE. The WINE data sets include provenance information, such as when an attack was first observed, where it has spread, and how it was detected. Moreover, experimentation is performed in a controlled environment at Symantec Research Labs and all intermediate and final results are kept within the administrative control of the system.

However, recording all of the mechanical steps in an experiment is not enough. To reproduce a researcher's conclusions, the hypothesis and the reasoning behind each experimental step must be explicit. To achieve this transparency, experimenters are provided with an electronic lab book (in the form of a wiki) for documenting all of the experimental procedures. Maintaining the lab book requires a conscious effort from the experimenters and produces documentation on the reference data sets created for the purposes of the experiment, the script that executes the experimental procedure, and the output data. Keeping such a lab book is a common practice in other experimental fields, such as applied physics or experimental biology.

5.2 WINE Analysis Example: Determining the Duration of Zero-Day Attacks

A zero-day attack exploits one or more vulnerabilities that have not been disclosed publicly. Knowledge of such vulnerabilities enables cyber criminals to attack any target undetected, from Fortune 500 companies to millions of consumer PCs around the world. The WINE platform was used to measure the duration of 18 zero-day attacks by combining the binary reputation and anti-virus telemetry data sets and by analyzing field data collected on 11 million hosts worldwide (Bilge & Dumitras, 2012). These attacks lasted between 19 days and 30 months, with a median of 8 months and an average of approximately 10 months (Figure 6). Moreover, 60% of the vulnerabilities identified in this study had not been previously identified as exploited in zero-day attacks. This suggests that such attacks are more common than previously thought. These insights have important implications for future

security technologies because they focus attention on the attacks and vulnerabilities that matter most in the real world.

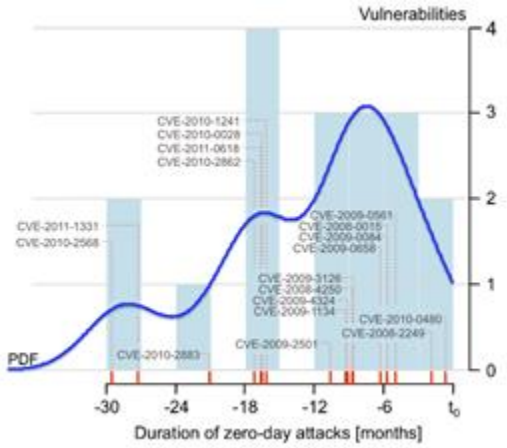


Figure 6. Analysis of zero-day attacks that go undetected.

The outcome of this analysis highlights the importance of Big Data techniques for security research. For more than a decade, the security community suspected that zero-day attacks are undetected for long periods of time, but past studies were unable to provide statistically significant evidence of this phenomenon. This is because zero-day attacks are rare events that are unlikely to be observed in honeypots or in lab experiments. For example, most of the zero-day attacks in the study showed up on fewer than 150 hosts out of the 11 million analyzed. Big Data platforms such as WINE provide unique insights about advanced cyber attacks and open up new avenues of research on next-generation security technologies.

6.0 Conclusions

The goal of Big Data analytics for security is to obtain actionable intelligence in real time. Although Big Data analytics have significant promise, there are a number of challenges that must be overcome to realize its true potential. The following are only some of the questions that need to be addressed:

1. Data provenance: authenticity and integrity of data used for analytics. As Big Data expands the sources of data it can use, the trustworthiness of each data source needs to be verified and the inclusion of ideas such as adversarial machine learning must be explored in order to identify maliciously inserted data.
2. Privacy: we need regulatory incentives and technical mechanisms to minimize the amount of inferences that Big Data users can make. CSA has a group dedicated to privacy in Big Data and has liaisons with NIST's Big Data working group on security and privacy. We plan to produce new guidelines and white papers exploring the technical means and the best principles for minimizing privacy invasions arising from Big Data analytics.

3. Securing Big Data stores: this document focused on using Big Data for security, but the other side of the coin is the security of Big Data. CSA has produced documents on security in Cloud Computing and also has working groups focusing on identifying the best practices for securing Big Data.
4. Human-computer interaction: Big Data might facilitate the analysis of diverse sources of data, but a human analyst still has to interpret any result. Compared to the technical mechanisms developed for efficient computation and storage, the human-computer interaction with Big Data has received less attention and this is an area that needs to grow. A good first step in this direction is the use of visualization tools to help analysts understand the data of their systems.

We hope that this initial report on Big Data security analytics outlines some of the fundamental differences from traditional analytics and highlights possible research directions in Big Data security.

Bibliography

Alperovitch, D. (2011). Revealed: Operation Shady RAT. Santa Clara, CA: McAfee.

Bilge, L. & T. Dumitras. (2012, October) Before We Knew It: An empirical study of zero-day attacks in the real world. Paper presented at the ACM Conference on Computer and Communications Security (CCS), Raleigh, NC.

Bryant, R., R. Katz & E. Lazowska. (2008). Big-Data Computing: Creating revolutionary breakthroughs in commerce, science and society. Washington, DC: Computing Community Consortium.

Camp, J. (2009). Data for Cybersecurity Research: Process and "whish list". Retrieved July 15, 2013, from http://www.gtisc.gatech.edu/files_nsf10/data-wishlist.pdf.

Cugoala, G. & Margara, A. (2012). Processing Flows of Information: From Data Stream to Complex Event Processing. ACM Computing Surveys 44, no. 3:15.

Curry, S. et al. (2011). RSA Security Brief: Mobilizing intelligent security operations for Advanced Persistent Threats. Retrieved July 15, 2013, from http://www.rsa.com/innovation/docs/11313_APT_BRF_0211.pdf

Dumitras, T. & P. Efsathopoulos. (2012, May). The Provenance of WINE. Paper presented at the European Dependable Computing Conference (EDCC), Sibiu, Romania.

Dumitras, T. & D. Shou. (2011, April). Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Enviornment (WINE). Paper presented at the EuroSys BADGERS Workshop, Salzburg, Austria.

Giura, P & W. Wang. (2012) Using Large Scale Distributed Computing to Unveil Advanced Persistent Threats. New York, NY: AT&T Security Research Center.

François, J. et al. (2011, May). Bottrack: Tracking botnets using netflow and pagerank. Paper presented at the IFIP/TC6 Networking Conference, Valencia, Spain.

François, J. et al. (2011, November). BotCloud: Detecting botnets using MapReduce. Paper presented at the Workshop on Information Forensics and Security, Foz do Iguaçu, Brazil.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Stamford, CT: META Group.

Luckham, D. (2002). The Power of Events. Vol. 204. Addison-Wesley.

Manadhata, P.K., W. Horne, & P. Rao. (Forthcoming). Big Data for Security: Processing Very Large Enterprise Event Datasets. In B. Furht and A. Escalante (Eds.), Handbook of Big Data Analytics. s.l: Springer.

Verizon Inc. (2010). 2010 Data Breach Investigation Report. Retrieved July 15th, 2013, from http://www.verizonenterprise.com/resources/reports/rp_2010-data-breach-report_en_xg.pdf

Yen, T.-F., Oprea, A., Onarlioglu, K., Leetham, T., Robertson W., Juels, A., Kirda, E. (2013, December) Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks. In Proceedings of the Annual Computer Security Applications Conference (ACSAC), New Orleans, LA.

