BIG DATA ANALYTICS FOR CYBER SECURITY

Bharath Krishnappa
Principal Software Engineer
EMC, India Center of Excellence
Bharath.Krishnappa@rsa.com

EMC²

Table of Contents

Introduction

According to a 2011 EMC-sponsored IDC study "Extracting Value from Chaos", it was estimated that 1.8 zettabytes will be created in 2011 and the world's information will double every year. A zettabyte is roughly 1000 exabytes. To place that volume in more practical terms, an exabyte alone has the capacity to hold over 36,000 years' worth of HD quality video…or stream the entire Netflix catalog more than 3,000 times [Cisco Blog]. Overwhelming, isn't it?

Today, only a fraction of this data is used to aid business decisions. Big data analytics is set to change this. Its goal is to bring more and more data into play. Many organizations and big businesses have woken up to the potential of big data and machine learning. Almost all of them are either exploring ways to put them to use or are already using them. Big data has been one of the top trending topics in Information Technology for some time now and I believe it will continue to be on top for a long time.

Cyber security is an ever evolving field. Every new technology will introduce a new set of threats and vulnerabilities, making security a moving target. What makes it worse is the fact that security is almost always an afterthought when it comes to new technologies. According to a news release from Ernst & Young - "Global companies in a hurry to adopt new technologies and media, leaving threats to security as an afterthought".

Considering the dynamic nature of the security domain, big data analytics can play a major role in areas such as malware detection, intrusion detection, multi-factor authentication, etc. Most organizations today tend to over-compensate with techniques such as multi-factor authentication to protect themselves and their customers. Security almost always trades-off usability. If not moreso, usability is almost as important as security for some verticals like ecommerce. For an ecommerce site each extra step or extra second required to complete a transaction will negatively impact revenue. Big data and machine learning techniques can be employed to assess risk by collecting and analyzing various contributing factors such as IP address, device type, device location, browser, MAC address, ISP, user history, etc. Only if the risk is high will additional security measures be enforced. This way, usability will be impacted only for a few transactions that are deemed risky. This article documents and discusses such examples where big data analytics techniques can be used to tackle some of the difficult security challenges like Advanced Persistent Threat (APT), big ticket breaches plaguing both private and public sectors today.

Big Data Analytics

This is how Gartner defines big data - “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” There are plenty of definitions out there for big data but Gartner’s definition made best sense to me. One notable omission in this definition is visualization. The reason I feel visualization is important is that the collection of data always outpaces our ability to derive value from it. This is due to limited availability of actionable insights. However, with the aid of visual analytics, the structures and patterns in the data can be analyzed and actionable insights can be derived at a rapid pace [Ngrain article].

Most techniques like machine learning, statistics, predictive analysis, behavioral analysis, etc. used today for big data analytics have been there for a long time. Traditionally, these techniques were used on structured data sets ranging from a few MBs to a few GBs. Today they can handle bigger volumes, up to petabytes of both structured and unstructured data. Drivers for this rapid change are:

- Hadoop and tools built around Hadoop that can handle all three Vs of big data; volume, velocity and variety.
- NoSQL databases that can store massive amounts of data and retrieve them at breakneck speed.
- Decreasing cost of storage and compute.
- Cloud computing technologies that enable easy and elastic access to massive amounts of compute, storage, and network.

Security

Analytics is not new to the world of security. If you think about it, intrusion and fraud detection systems have been using analytics for a long time. But these traditional systems employ analytics in a limited way.

- They collect very limited data due to storage or relational database management system (RDBMS) restrictions.
- Data is deleted after a fixed retention period because of storage or RDBMS capacity or performance restrictions.
- The amount of processing involved for decision making is also limited since they are expected to be non-intrusive and add limited overhead.

- To meet non-intrusive (high tps, low response time) requirements, these expensive techniques are employed as periodic offline process, delaying detection. Needless to explain the cost of delays in detecting fraud and intrusion.
- They can handle only structured data.

If you look at the limitations of these traditional systems and the reasons for the restrictions, the restrictions seem artificial when you factor in big data technologies. Hadoop and NoSQL databases can:

- be deployed on commodity hardware and handle massive amounts of data
- scale horizontally; as and when the data size increases, more nodes can be added seamlessly
- run complex processes on massive amounts of data at unprecedented speeds.

Add real time stream processing to the mix and we can build systems with limitless capabilities. In this section, we discuss some of the limitations of the traditional security solutions and how big data tools can be utilized to alleviate them.

Intrusion detection

Intrusion detection systems (IDSs) monitor network, network node, or host traffic and flag any intrusions. IDS use either statistical anomaly-based technique or signature-based technique for intrusion detection.

Signature-based techniques monitors and compares the network packets and traffic patterns against a set of signatures created from known threats and exploits. Big data tools may not add a lot of value for this technique except that they can improve the pattern matching speeds and increase the capacity of signature databases.

On the other hand, statistical anomaly-based detection technique compares the network traffic with the baseline and flags major deviations from the baseline as intrusion. In my opinion, this is the better technique compared to signature-based technique for the simple reason that it is adaptive. Here big data tools can play a huge role. It can facilitate collection and extended retention of more data per network packet and also the traffic pattern and monitoring information. This anomaly-based technique is as good as the baseline with which it compares. If the baseline has to be effective, it has to be dynamic and contextual. The baseline can differ

based on the time, events, etc. Below are a few examples to demonstrate the contextual nature of baseline.

- In organizations operational mostly during the day, network traffic is less at night.
- Traffic to the ACH handling module increases during the time of day when ACH transactions are processed.
- There should be no network traffic in the daytime from the laptop belonging to the user who works the night shift.
- Traffic to a bank on the last day of the month can surge when salaries get credited.

Most of the information required to build the contextual baseline like employee shift timings, ACH processing window, etc. are already available in different systems but they are not leveraged yet. Big data techniques like real time stream processing can leverage data collected from such disparate sources and build contextual baselines at great speed. Add some guided or semi-guided learning techniques to the mix and utility of the IDS will improve significantly. It can reduce the false alarm rates that plague them today and even when there are false alarms, it can learn some specifics from the analyst and use that to fine tune the baseline. It can easily combat IDS evasion techniques like low bandwidth attacks, fragmentation, etc.

Remote Banking Fraud detection

Banks have a very difficult task of balancing between fraud detection overheads and the ever increasing need for faster payments. Since faster payments translate to instant revenues, it almost always trumps fraud detection. Due to this, the fraud detection solutions in this space have to be very accurate. Tolerance for false positives is very low. A list of factors that can help determine frauds accurately are:

- User-to-device mapping – Maintain a list of devices user employ to access their account.
- Device profile – If the device is a public access computer, no anti-virus protection, etc.
- Source IP address – Is the IP blacklisted or is it a proxy IP?
- Location of access – i.e. if the access is from countries like Nigeria
- User behavior – time of day, last access location
- Payee profile – is the payee of a transfer known to have previously benefited from fraud?
- User's risk appetite – Credit rating, casino statistics, traffic violations, law violation history. The thought behind listing them as factors is that a risky transaction from a person who is accustomed to taking risks may not really be a fraudulent transaction.

- User's profession and education levels – A security professional is less likely to open a phishing email compared to a fashion model.
- Does the user have a history of substance abuse? –irrational choices can be made when under the influence of substances.
- Is the user a public figure? – high value target.
- User's travel itinerary from travel sites.

I was letting my imagination run amok; some of these facts cannot be collected in a secure and trustworthy fashion. Other facts are out of bounds for the banks due to laws and regulations. However, suppose there is a framework for organizations to share data about the users and devices. Imagine how much the following examples can contribute to the accuracy of fraud detection:

- a user's travel itinerary from travel sites like Agoda and Expedia
- type of sites that user frequents that Facebook and Google tracks,
- devices on which anti-virus products are installed from the anti-virus vendors

Solutions available in this space today already use some of these factors. But when it comes to collecting more data, they are limited by capacity constraints of the traditional technologies. Even the data retention intervals are short due to capacity constraints. In these solutions, building of profiles is handled by offline scheduled tasks to limit the overhead of fraud detection but this increases the time to value of the data. An analyst's turnaround time is typically on the higher side because mining for the required data is a slow and tedious process. NoSQL databases can alleviate capacity constraints to a large extent and It ease the pain of data mining for fraud analysis. Tools like Storm that offer distributed real time computation capabilities can be used to build the profiles quickly and on the fly to reduce the time to value of data. Big data visual analytics aids for fraud analysts can help them derive new actionable insights and push them into the system.

Monitoring, Breach investigation, and Incident response

An RSA white paper – Intelligence Driven Threat Detection & Response – explains the limitations of the traditional monitoring techniques very well - "Faced with constant streams of data moving over the network, the temptation—and until recently, the only option—has been to focus data collection and analytics on select problem areas. Security teams often collect and analyze logs from critical systems, but this log-oriented approach to threat detection leaves

many blind spots that sophisticated adversaries can exploit." The traditional techniques also relied heavily on signature-based detection and black listing. These techniques can be bypassed and when they are it takes significant time to block the new attack vector because these techniques are by nature, static.

The traditional techniques are also weak at identifying sophisticated attacks like APT and steganography.

APT - CSA's "Big Data Analytics for Security Intelligence" paper defines APT as "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."

Steganography is a technique of avoiding detection by hiding information in images or other media files that are usually considered non-risky and are subjected to limited monitoring.

An intelligence-based approach to monitoring with the aid of Big Data technologies can address all these limitations of traditional systems. To start with, not having to be concerned with capacity constraints, the monitoring tools can start gathering all the network packets, logs, etc. instead of focusing only on the critical and problem areas. It can start engaging deeper and more complex packet inspection and log analysis techniques by leveraging the scalable parallel processing big data techniques. Visual analytics can be used to provide comprehensive network visibility to the network security administrator. It can even focus or highlight areas that are deviating from the usual pattern and facilitate quick drill down and rollup functionality that will aid in faster identification of threats. Additionally, it could spot stealthy techniques like APT by identifying many minor deviations or intrusions from the same user or device, weaving them together and flagging them as a whole.

In a blog post, Bruce Schneier states, "Security is a combination of protection, detection, and response." In the same blog post he stresses the importance of incident response and how speed is off essencial when it comes to incident response. The ability to see all the alerts in a centralized security management console and drill down capabilities to quickly wade through the specifics would help in this regard. The ability to quickly re-construct and view the activities of

the impacted and impacting systems can accelerate breach investigations and identify other systems in network that are similarly impacted. The only way to build such capabilities is by adopting technologies like NoSQL for faster data retrieval and visual analytics to facilitate them to quickly drill down to the problem area and rollup to see if other network areas are similarly impacted.

Usability and convenience

Security controls are almost always a trade-off between usability and convenience. To keep a relative few at bay, all of us are expected to make seemingly small sacrifices every day. For example:

- The multiple times that we are expected to enter passwords throughout the day.
- At times applications expects us to log in with more than one credential (multi-factor, step-up authentication) to augment security.
- The long queues in airports, malls, and other public places for security clearance.
- The procedures employed by call centers to ensure your authenticity before they actually resolve or address your concerns/complaints. Not only are these procedures frustrating for us because we need to spend a lot of time on the phone even to get minor clarifications, this is a significant cost for the call centers too.

These controls are in place because of the inefficiencies and limitations of current technologies to determine risk accurately. The multiple login problems can be addressed by calculating the risk based on some of the factors I have listed in the “Remote Banking Fraud detection” section and stepping up the security controls based on risk. Couple this technique with SSO technologies to ensure that we don’t see often boring and at times frustrating login screens.

At airports, big data tools and technologies can be leveraged to build risk profile of a person in real time and make it available to security officers. Risk profiles can be built by pulling already available data from various sources and could also factor in vitals like blood pressure, heart rate, and any variation in them while approaching the turnstile or the security officer, etc.

Challenges

Although big data technologies hold the key for most of our problems and for future innovations, there are plenty of challenges that have to be addressed quickly. The biggest challenge and concern is that big data technologies can erode privacy. The greatest strength of big data technologies is its ability to locate a needle in a haystack. This capability can be used to easily

dig into data about people that they intend to keep private. Suppose a person wants to keep his age private and decides to hide his date of birth on all of his social profiles. A person looking for his age can still easily determine it by looking at the profiles of people who are listed as his classmates. While this example is just a minor privacy infringement, if you understand the capabilities of NORA (Non Obvious Relationship Awareness) systems, you will know the full extent of privacy erosion that such technologies can cause when coupled with big data technologies and data harvested for big data analytics. Other challenges like weak security controls of big data tools, data provenance, etc. - though major challenges in their own right - pale in comparison with privacy challenges. In this section, we discuss some of these challenges in detail.

Privacy

In the interest of keeping the arguments balanced, I will first touch upon the virtues of big data and the role it will play in the evolution of human race. I will then discuss the privacy erosion that can be caused by big data tools and data harvested for big data analytics.

Data is raw and insights are derived by analyzing patterns and structures in the data. The insights of today become data points of tomorrow from which new insights can be gained. For example, Newton's laws of motion were insights when he first came up with it, but today it is a data point. These iterations of data to insights and insights becoming data points for the next iterations are drivers for technological progress. Newton expressed a similar thought very well when he said – "If I have seen further it is by standing on the shoulders of giants." Big data technologies are set to accelerate these iterations that drive technology advancements. Even though most traditional analytical approaches were good at mining data and finding answers to specific questions, big data technologies can lead to unexpected insights that were not even being sought in the first place. It is this groundbreaking capability of big data technology that helped Walmart figure out that there is usually a lot of demand for Strawberry Pop Tarts before a hurricane. If not for these techniques, what were the chances of discovering this insight? It is this remarkable capability of big data technologies that sets it apart and can drive radical/disruptive innovations even in fields like automobile engineering which hasn't seen any radical innovation in decades. Taken further, in the field of medicine, it could pull a lot of non-obvious but high sensitivity and specificity biomarkers.

Clearly, an aspect of big data technologies that many of us fear is that it can erode personal privacy. Is this fear unfounded?

The answer is a most definite no. Today most of us are monitored in many different ways for the purpose of deterring crime, maintaining law and order, etc. For example, we find cameras in most public places like ATMs, cinemas, malls, and airports. Until recently this data was not networked together. However, the lure of the value proposition that can be gained though big data technologies is undeniable. Once networked together, big data tools make it possible to unearth information about individuals that can be easily misused [Bryant, Katz, & Lazowska, 2008].

Almost all new age Internet companies consume a lot of user data. Websites track our activities on their sites but they do not stop there; they track us even when we are on other sites. Such companies will discover and already know a lot about us. Sometimes they know things about us of which even we are not aware. Such data gives them a lot of power over us. Everyone has skeletons in their closet and such data can be used to know which closets hold what skeletons. If the people at the helm in such organizations decide to misuse the data can even bring heads of state to their knees. This is a lot of power. And, as the saying goes, “Absolute power corrupts absolutely.”

If you have heard about the Sony hack, there is one big take away for all the cyber criminals. Governments and heads of state cannot be blackmailed because most countries have strict non-negotiation policy with terrorists, criminals, blackmailers, etc. But private companies can give in to blackmail very easily because in the private sector, decisions are typically made based on cost analysis.

This is why I feel we should not let the private sector self-regulate in this aspect. Governments should play an active role in framing new policies and regulations on how data should be safeguarded. To weather the rapid changes in technology, policies and regulations should not be stated in terms of tailored solutions for different technology tracks, and sub-tracks. Instead, they should be stated in terms of intended outcomes [PCAST].

Governments should also set right their own houses. It is well-known that governments track a lot of our activity and they themselves do not have comprehensive checks and controls to prevent misuse. There are instances where government employees have misused this data to spy on their friends. Governments and government agencies are going against the basic privacy principle of using data only for the purposes that it was collected by forcing firms to share data

about their users, partners, etc. and coaxing security firms to build backdoors in security solution to facilitate surveillance. [Manadhata, Horne, & Rao]

Widespread tracking and surveillance by governments and private organizations without respect for geographical and jurisdictional boundaries is giving rise to self-censorship all over the world. Governments would be wise to start rebuilding their credibility by bringing in regulations to safeguard the surveillance data that they collect and strict laws that criminalize the misuse of such data.

Security controls in big data tools

Thus far the discussion has focused on how big data can be used for security, but good security controls for big data tools is a necessity if these tools are to be leveraged for security purposes. While big data tools may far outdo traditional data tools in terms of data retrieval and analysis speeds, when it comes to security models, traditional tools have the upper hand by far. The authentication and authorization mechanism employed in most of these tools are lax and inefficient. The RBAC controls that they provide are not granular enough. The options to secure inter-node communications are very basic. Since these tools have unlimited capacity to consume data, they also need stricter controls to prevent unauthorized access. Unfortunately, the security controls and options that are available for these tools are grossly inadequate to address the risks associated with the massive scale of data.

Even data protection solutions like backup and recovery are not very comprehensive and lack options when it comes to big data tools. These gaps should be quickly plugged if big data technologies have to be used for security purposes. If not, it will become a very serious and obvious handicap.

Data Authenticity and Integrity

The quality of analytics results greatly depends on the quality of data that it analyzes. Inferior data quality not only impacts the immediate results, it can impact the models in ways that could impact the quality of the future results as well. For security analytics, the data quality attributes are authenticity and integrity. Techniques like log signing should be adopted to ensure authenticity and integrity of logs. Similar techniques should be adopted to ensure trustworthiness of other data used for analytics. Also, defensive techniques should be built to catch attempts to insert noise into input data to distort results and bypass security controls.

Conclusion

Big data holds enormous potential. It will definitely be the engine that drives unprecedented rates of innovation in all fields. Like other industries, it has the capacity to transform the field of security too. Only with the use of big data technologies can we combat the kind of threats and attacks we are experiencing today. But the job of protecting against attacks does not stop there. Adversaries are also going to leverage the power of these technologies to build more sophisticated, stealthy, and complex attacks. Many of us acknowledge that attackers are often smarter and better organized than the security industry and private sector organizations. The lack of jurisdictional boundaries in the Internet and the difficulty of tracing an attack back to an individual makes them even stronger. Even when the attack can be traced back to an individual, it is difficult to identify and collect irrefutable evidence because factors like source-IP address, geo-location, etc. that help trace attackers are spoofable.

One of the six laws of Melvin Kranzberg states, “Technology is neither good nor bad; nor is it neutral”. This is true even for big data technologies. Big data technology is already here and most of us are aware of its big value proposition. I believe most of the debate today is whether we should allow or stop collecting certain types of data that could invade privacy. However, I think at this juncture it makes more sense to debate about regulations, techniques, and solutions to solve this challenge without reducing the efficacy of big data technologies. If big data technologies have to be leveraged effectively for security solutions, there should be more focus on building effective and granular security controls for the tools and also on solutions to ensure trustworthiness of the input data.

References

[Cisco Blog] Thomas Barnett, Jr. “The Dawn of the Zettabyte Era [INFOGRAPHIC]”, June 23, 2011, retrieved from http://blogs.cisco.com/news/the-dawn-of-the-zettabyte-era-infographic on 14-Jan-2015

[RSA white paper] RSA “INTELLIGENCE DRIVEN THREAT DETECTION & RESPONSE”, retrieved from https://www.emc.com/collateral/white-paper/h1304-intelligence-driven-threat-detection-response-wp.pdf on 12-Jan-2015

[Ngrain article] Ngrain “3 reasons why “visualization” is the biggest “V” for big data”, retrieved from http://blogs.cisco.com/news/the-dawn-of-the-zettabyte-era-infographic on 14-Jan-2015

[CSA] CLOUD SECURITY ALLIANCE. “Big Data Analytics for Security Intelligence”, September 2013, retrieved from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf on 5-Jan-2015

[Bryant, Katz, & Lazowska, 2008] Bryant, R., R. Katz & E. Lazowska. “Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society”, December 2008 , retrieved from http://www.cra.org/ccc/files/docs/init/Big_Data.pdf on 3-Jan-2015

[Manadhata, Horne, & Rao] Manadhata, P.K., W. Horne, & P. Rao. “Big Data Analytics for Security”, retrieved from http://www.utdallas.edu/~alvaro.cardenas/papers/IEEESnP.pdf on 5-Jan-2015

[PCAST] President’s Council of Advisors on Science and Technology. “REPORT TO THE PRESIDENT BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE”, May 2014, retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf on 7-Jan-2015