White Paper

HPCC Systems for Cyber Security Analytics

See through patterns, hidden relationships and networks to find threats in Big Data.

September 2012

LexisNexis®

Risk Solutions
Government

HPCC Systems for Cyber Security Analytics

Many of the most daunting challenges in today's cyber security world stem from a constant and overwhelming flow of raw network data. The volume, variety, and velocity at which this raw data is created and transmitted across networks is staggering; so staggering in fact, that the vast majority of data is typically regarded as background noise, often discarded or ignored, and thus stripped of the immense potential value that could be realized through proper analysis. When an organization is capable of comprehending this data in its totality—whether it originates from firewall logs, IDS alerts, server event logs, or other sources—then it can begin to identify and trace the markers, clues, and clusters of activity that represent threatening behavior.

Today's biggest cyber challenges, which include the emergence of the advanced persistent threat, take advantage of the data deluge described above to establish long-term footholds, exploit multiple vulnerabilities, and deliver malicious payloads, all while avoiding detection. This white paper will focus on the big data processing platform from LexisNexis called HPCC Systems, (High Performance Computing Cluster) as a technology platform to ingest and analyze massive data that can offer meaningful indicators and warnings of malicious intent.

In contrast to current approaches, the effectiveness of the HPCC Systems solution *increases* as data volumes grow into the hundreds of terabytes to petabyte range. Not only does this solution provide the ability to fuse an organization's own network data (e.g. firewall logs, access logs, IDS alert logs, etc.), but also it delivers enrichment routines that can automatically incorporate and fuse any relevant 3rd party data set, including blacklists, known bad domains, geo-location data, etc. Finally, HPCC Systems delivers this capability at speeds that cannot be achieved by typical database-oriented approaches. The results of HPCC Systems large-scale analytics can provide an administrator with a significant forensic advantage and a tremendous head start in quickly verifying the significance of a potential incident.

The Advanced Persistent Threat

Consider the following situation: the young man emerged from his supervisor's office with a job to do. This man - a systems administrator at a large laboratory in on the west coast - was asked to investigate a minuscule accounting error in his lab's computer usage accounts. He likely had more interesting things to do on a sunny August day than research a $0.75 discrepancy, but the more he pored over access logs and system files, the more curious – and suspicious – he became. As he followed the faint trail of electronic breadcrumbs, he realized that this seemingly benign "accounting error" was the result of intentionally malicious activity. It was, in fact, the first of many subtle clues left behind by a hacker who had gained access to the administrator's network by exploiting a vulnerability in the lab's email system.

Was this simply a prank? Could it have been a bored college student with too much free time and a mischievous streak? This seemingly minor incident turned out to be the start of a ten-month journey that would take the systems administrator from his network's data center all the way to the heart of several U.S. military networks. He would employ all his accumulated knowledge to monitor, analyze, and deploy various kinds of electronic bait for the hacker. Over time, he was able to discover not only the root of these intrusions, but the motivation behind them as well. He traced the attacks to a hacker living in central Europe.. Using a persistent, methodical approach – over a period of months – this hacker had not only gained access into the laboratory's network, but was able to exploit a number of interconnections between national labs, government agencies, and government contractors to gain root level access to military computers around the United States. Furthermore, the hacker used this access to download hundreds of sensitive documents related to nuclear weapons and defense programs. This hacker, as it turned out, had been systematically exploiting network vulnerabilities to access these facilities and was selling the information he had illegally obtained to agents of a foreign government.

While this tale reads like a cyber attack that might have been pulled from today's headlines, these events actually unfolded in a year remembered more for Chernobyl, the Iran-Contra affair, and the Space Shuttle Challenger disaster - 1986. This was an early form of a network attack now referred to as "Advanced Persistent Threat" (APT). Like this example, a typical modern day APT is characterized by:

- Attackers who are typically *funded and directed by external entities*, organizations and governments
- Attackers who utilize the *full spectrum of intelligence collection* methods, which may include computer intrusion technologies as well as coordinated human involvement and social engineering techniques to reach and compromise their target
- An attack that is conducted through *continuous monitoring* and interaction in order to achieve the defined objectives
- An attack that, rather than relying on a barrage of continuous intrusions and malware, employs a *"low-and-slow"* approach

As our world grows increasingly more connected, LexisNexis recognizes that Advanced Persistent Threats have the potential to cause increasingly significant damage to critical infrastructure, financial systems, and to sensitive military operations. Examples are frighteningly numerous, and include well-publicized incidents such as Stuxnet[1], Titan Rain[2], and Operation Aurora[3]. While the basic features of the threat are not much different than they were in 1986, the potential for damage has been greatly magnified.

Try as we might, these attacks are difficult to detect. Why? The attackers are careful, patient, well funded, and highly motivated. They tend to apply methodical techniques that keep their activities under the radar. Whereas an attack such as a distributed denial of service is generally hard to miss; the probes, connections, and malicious downloads performed by sophisticated actors over the course of months or years are easily obscured by huge volumes of routine network activity.

Therein lies the crux of the problem. With the rise of mobile computing, distributed data storage, cloud infrastructures, and Internet-enabled telecommuting, we are witnessing three phenomena which provide attackers tremendous opportunities to do harm:

1. Multimedia, networked collaboration, and mass participation have resulted in a constant deluge of heterogeneous network activity, both within private networks and across the public Internet. ***This provides constant "cover" to the activities of malicious users.***
2. Through technologies like connected mobile devices, virtual private networks, and cloud computing, the notion of a corporate network has expanded well beyond the traditional firewall-based perimeter. ***This provides significantly more vulnerabilities and access points through which malicious users can gain entry to protected resources.***
3. The combination of increased activity and expanded network architectures has complicated network security. Maintaining a secure posture is a result of constant vigilance and mitigating one's risk through vulnerability management and continuous monitoring. However, security technologies are unable to keep pace with the evolving threat landscape, and as a result traditional approaches have shown severe limitations. ***These limitations, and the ability to overcome them, are the focus of this white paper.***

LexisNexis HPCC Systems for Deep Forensics Analysis

HPCC Systems is a massively parallel analytics platform that delivers two large-scale, long-term data fusion capabilities. When applied to the cyber security domain, HPCC Systems provides network security teams the ability to transform massive data to intelligence in a manner that would be impossible for traditional data mining and analysis technologies. The HPCC Systems technology is optimized for aggregating, fusing, and analyzing massive, multi-source, multi-format data sets. It delivers an analytics capability that bridges the data gap between today's short-term operational data analysis and the deep situational understanding that only comes with large-scale data analytics.

The two core capabilities of the HPCC Systems include:

1. **Pre-Computed Analytics** – combined with continuous monitoring tools such as IDS and SIEM, pre-computed analytics improve the quality and accuracy of alerts by instantly comparing alert metadata against a comprehensive repository of historical network patterns and computed behaviors. This adds relevance and context to real-time alerts to not only identify and tag false positives, but to also associate seemingly benign activity with longer term, more serious threats, thereby addressing the "false negative" dilemma.
2. **Deep Forensics Analysis** – using an advanced query engine, security analysts with deep domain expertise and technical skills can routinely perform highly customized, sophisticated analysis as needed - against massively complex data sets. For instance, an analyst might want to execute a complex correlation across months' worth of log files originating from dozens of device types. For a typically large network, not only would this analysis need to fuse data of multiple formats, but it would also be required to correlate potentially hundreds of Terabytes (or more) of raw data. HPCC Systems delivers this capability at speeds that cannot be achieved by typical database-oriented approaches.

Combined, these capabilities are meant to tackle the challenge of "big data" in order to eliminate the cover attackers rely on and to provide better visibility into the increasing number of network entry points and vulnerabilities.

Pre-Computed Analytics for Cyber Security

There is a compelling need for new types of analytics, focused on massive, long-term data sets. The federal government is pressing its agencies for "continuous monitoring" of government networks. However, traditional approaches to continuous monitoring are limited by the amount of data they can analyze. As a result, systems such as Intrusion Detection Systems are restricted to performing "selective analysis" – either looking at some metadata subset (selected fields from packet headers or netflow messages, for instance), or sampling network data to seek statistically significant patterns (analyze one of every thousand packets). Additionally, signature-based detection algorithms typically analyze data over relatively small time windows (hours or days), and so tend only to be able to detect short-term activities. In contrast to these technologies, LexisNexis approach to automated information analysis:

- Processes *all* data, regardless of the overall volume.
- Merges data from *many sources*, whether they are structured data sets, unstructured text, or feeds from external sources.
- Performs *full-text analysis* on *all fields* in the data
- Executes *large scale analysis in a timely fashion*, thanks to a massively parallelized data processing technology

HPCC Systems delivers a default library of configurable input adapters and pre-computed cyber security analytics. The computed results of these analytics can be called by end users through a web-based search interface, or automatically queried by 3rd party solutions such as Intrusion Detection Systems (IDS) and Security Information Event Managers (SIEM). They are "pre-computed" in that analytic results are routinely calculated over all available data, and prepared for consumption by end users or third party systems.

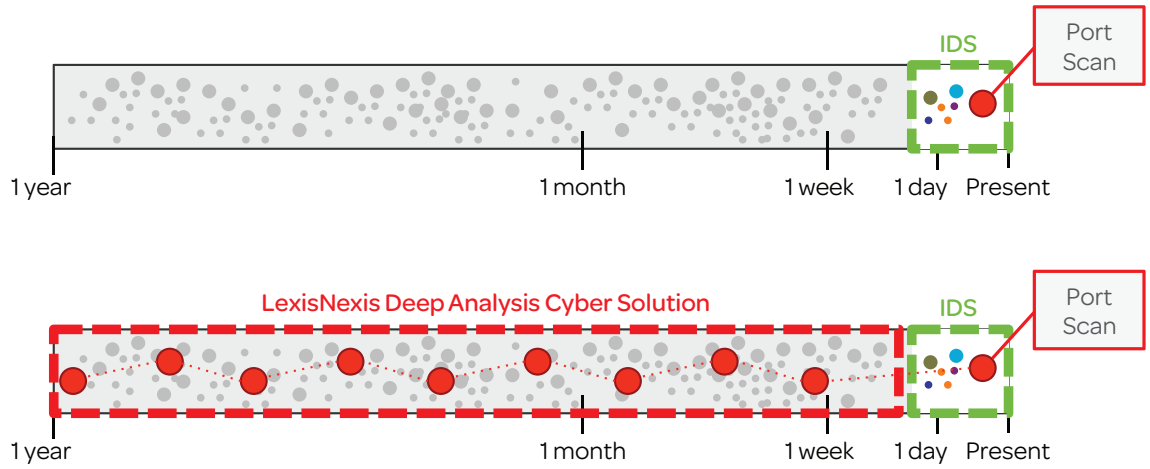

Figure 1: Continuous monitoring technologies alone cannot perform long-term contextual analysis. When deployed in conjunction with the Deep Analysis Cyber Solution, new alerts from these technologies can be linked to historical behavior and attack patterns to identify persistent, "low and slow" attacks.

The solution works by routinely analyzing *all* collected data from relevant sources, and then computing a series of analytical results. While real-time network data is fed to continuous monitoring systems via techniques such as packet capture; the HPCC Systems can routinely operate on aggregated log output from network devices (firewalls, routers, servers, etc.) and security systems (vulnerability management, NIDS, HIDS, antivirus, etc.). When this aggregated output is ingested, HPCC Systems can fuse all data, and re-computes a series of cyber analytics against either the entire set of data, or just incremental portions. Once the analytics have been applied to the target data, computed results are persisted, indexed and prepared for delivery via standard web services or web-based interfaces.

From a forensics perspective, this query library comprises a comprehensive set of long term patterns that a cyber security operator would want to identify as part of an investigation into any potential exploit or alert.

Unlike typical analytical approaches, these queries are asked at scale, across a period of months and against tremendous volumes of data. The combined results of these queries can provide an administrator with a significant forensic advantage and a tremendous head start in concluding the significance of a potential incident.

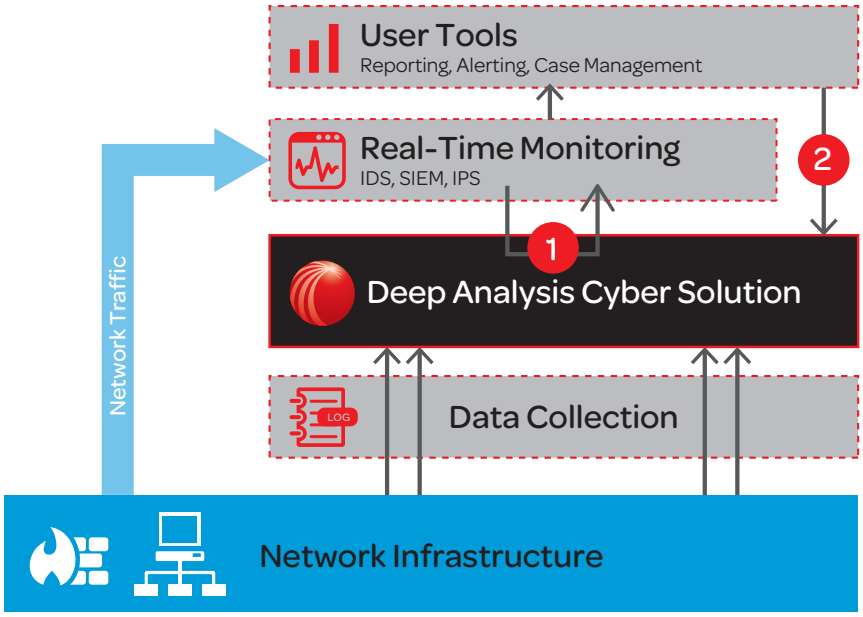

Figure 2: While real-time monitoring tools continuously analyze network traffic, the Deep Analysis Cyber Solution collects raw data, and routinely applies a series of long-term analytics. The results of these queries are made available to 1) continuous monitoring tools to enrich and validate alerts, and 2) end users via web-based searches or other applications.

Some of these default queries include:

Wavelet transforms. This mathematically derived set of functions was developed to analyze data across both frequency and temporal scales. This type of algorithm is especially useful for recovering a "true signal" from very noisy data sets, without requiring prior knowledge of anomalous patterns or needing explicit signatures to be defined. Standard wavelet algorithms can be effectively distributed across the Deep Analysis Cyber Solution's distributed computing cluster, and therefore can be applied to uncover a "true signal" within tens or hundreds of TBs of network "noise".

Information theory. This is a technique for distinguishing malicious network scan traffic from normal activity using general data compression tools. This method is particularly effective when a malicious user employs techniques that cause repetitive communication patterns, such as port scans occurring over large time periods utilizing multiple IP addresses to obfuscate malicious activities.

Slow oscillators. This query identifies groups of suspicious hosts by detecting communications patterns characterized by a small number of packets being sent between two hosts, or infrequent and irregular communications intervals.

Low hitters. This query identifies groups of suspicious hosts by detecting small numbers of systems in a large network that are suddenly making large quantities of DNS queries for new/previously unseen DNS entries. This can be representative of compromised machines trying to communicate back to command and control nodes that utilize "fast flux" techniques to constantly change the IP address.

Network activity clustering. Clustering can be used to find groups of features in the data. These might be IP addresses that demonstrate similar traffic patterns such as beacons, botnets and Trojans or groups of email servers. Various clustering techniques are employed, including K-means and leader clustering for processing IP packet data and NetFlow data.

Data exfiltration. This query identifies hosts, for any given time period, that are transmitting above or below a bytes per packet threshold for traffic on a given port.

Session hijacking detection. This query analyzes activity over every network connection to identify anomalies, such as a changing user-agent string, that are indicative of session hijacking activity.

Rare targets. This query seeks out long-term activities that are limited to a small number of hosts. For instance, it may look for external domains that are targeted by fewer than 10 different IP addresses, or oscillators that have targeted only rare domains.

Given a combination of search parameters (e.g. source IP, destination IP, source port, etc.), these queries can be executed and the results merged with the output of a traditional alerting system, such as IDS. This enhanced intelligence layer is used to enrich the IDS alert and to more effectively validate and prioritize it.

The Benefits of Pre-Computed Analytics

This capability improves an organization's continuous monitoring capability by allowing security administrators to relax tuning constraints on their alerting systems. Today, there is a tendency to over-tune continuous monitoring tools, like IDS, to reduce alert volumes to a manageable amount. The consequence of this, however, is that the monitoring system is then only able to detect major violations and blatant exploits, allowing nearly any action that is part of an APT to go unnoticed. By relaxing IDS rules to trigger on more events, and relying on deep, historical context to prioritize the resulting alerts, this solution delivers a two-fold benefit: first, deep historical context helps eliminate false positives so security administrators are left with a manageable workload, and more importantly, it reduces the number of false negatives, or malicious activity which would otherwise have gone unnoticed by an over-tuned continuous monitoring system.

Deep Forensics Analysis

While the pre-computed analytical routines of HPCC Systems represent an important capability, the key to effectively detecting and responding to Advanced Persistent Threats is in persistent vigilance and continuous human analysis. In fact, there are many cases when certain indicators and warnings push security experts into action, and when those experts are required to perform unique, "one off" analyses to make sense of malicious behavior. In these cases, a security analyst is expected to behave more like a traditional intelligence analyst, performing very deep, multi-dimensional analysis of their data.

Today's data mining technologies do not support the kind of improvisational analysis that's required to fully comprehend the nature and extent of an attack. The reasons for this vary; from the rigid relational database models that restrict the kinds of queries an investigator can perform to the inability to execute such queries against massive amounts of data within a desired timeframe (e.g. being able to perform calculations in minutes or hours instead of days).

This is where HPCC Systems shines. The technology is exceptional at this type of ad hoc usage – it is currently leveraged in multiple Federal programs specifically to fuse and link disparate structured and unstructured datasets and discover non-obvious relationships and anomalous patterns across truly massive content sets. It allows subject matter experts, possessed of deep technical skills, to write any kind of query against multi-terabyte, multi-source data sets.

The solution's non-relational storage architecture frees analysts from the shackles of traditional database-oriented approaches, where the kinds of analysis they can perform are restricted by how the data was modeled, and which individual fields the data modeler determined needed to be searchable.

In addition to fully customizing and re-purposing any of the included analytics, cyber analysts can develop, share, and reuse highly complex, custom analyses such as:

- Graph analytics to identify "hubs" of activity within massive data sets
- Multi-watchlist analysis and fusion against all-source network data
- N2 analysis where, for instance, every internal IP address might be compared against every other internal IP address that connected to the same external host on the same day. This kind of analysis can be performed by the Deep Analysis Cyber Solution in a fraction of the time it could be performed on a traditional relational database system. In fact, this kind of query at this scale will typically fail on most relational database servers.

As new threats emerge, analysts leverage the data-oriented, declarative programming language of HPCC Systems to model new attack patterns. These routines are executed against any size data set across the system's distributed computing cluster – so the analyst can run a custom query against 1 week's worth of data, or 6 months worth of data without having to worry about how to optimize the query for the target data set.

Conclusion

As more and more critical systems and mission operations become interconnected, the nature of national security and corporate threats is quickly shifting. It seems increasingly likely that the next major terrorist attack will be launched through an infected network host or a compromised industrial process rather than a suicide bomber or explosives-laden truck. Likewise the nature of conflict itself is evolving, as armies of hackers continually attempt to penetrate our most sensitive networks to steal classified information, trade secrets, intellectual property; or worse. Attacks of this nature require time, resources, and a motivated party to successfully execute. HPCC Systems has been developed to provide cyber security experts one of the tools they need to uncover the markers of an Advanced Persistent Threat before it has the opportunity to achieve its intended objectives. The HPCC Systems platform provides the capability to perform sophisticated analysis against all data over a long period of time. This comprehensive analysis - where no potential clue is ignored - is what allows the detection of subtle activities and "low and slow" attack patterns. The combination of pre-computed analytics and an ad hoc deep forensics analysis capability allows cyber security teams to both improve the quality of their ongoing automated monitoring and to quickly react to and understand new threats in an improvisational fashion.

Sources

[1] Stuxnet is a Microsoft Windows computer worm discovered in July 2010 that targets industrial software and equipment. Source: http://en.wikipedia.org/wiki/Stuxnet

[2] Titan Rain was the U.S. government's designation given to a series of coordinated attacks on American computer systems since 2003. Source: http://en.wikipedia.org/wiki/Titan_Rain

[3] Operation Aurora is a cyber attack that originated in China, and occurred from mid-2009 through December 2009. The attack targeted dozens of major major corporations, including Google.
Source: http://en.wikipedia.org/wiki/Operation_Aurora

For more information:

Call 877.316.9669
or visit http://hpccsystems.com

About HPCC Systems®
HPCC Systems® from LexisNexis® Risk Solutions offers a proven, data-intensive supercomputing platform designed for the enterprise to process and deliver Big Data analytical problems. As an alternative to Hadoop and mainframes, HPCC Systems offers a consistent data-centric programming language, two processing platforms and a single architecture for efficient processing. Customers, such as financial institutions, insurance carriers, insurance companies, law enforcement agencies, federal government and other enterprise-class organizations leverage the HPCC Systems technology through LexisNexis® products and services. For more information, visit http://hpccsystems.com.

About LexisNexis® Risk Solutions
LexisNexis® Risk Solutions (www.lexisnexis.com/risk/) is a leader in providing essential information that helps customers across all industries and government predict, assess and manage risk. Combining cutting-edge technology, unique data and advanced scoring analytics, we provide products and services that address evolving client needs in the risk sector while upholding the highest standards of security and privacy. LexisNexis Risk Solutions is part of Reed Elsevier, a leading publisher and information provider that serves customers in more than 100 countries with more than 30,000 employees worldwide.