A Big Data Approach to Minimizing Cybersecurity Threats

Abstract

The rise of digitalization across industries has impacted the nature of workflows, thereby amplifying the velocity and sophistication of cyberattacks. To prevent such threats from increasing, it is critical to quickly contextualize, correlate information, detect, and predict anomalies in real-time.

The traditional tools to handle security threats that emerged at the turn of the century are no longer relevant. Today, data has become complex and is regularly generated in vast amounts from numerous sources. Securing such data wealth now requires powerful and advanced systems.

A specialized security data lake or a security information and event management (SIEM) platform atop a fully scalable big data analytics platform can handle such sophisticated data and their rising volumes. A scalable big data and machine learning platform along with real-time threat data can generate a robust threat detection model, which can then identify new cyberattacks.

This paper delves into the challenges in today's cybersecurity era and how a big data-based security data lake platform can enhance the functionalities of SIEM.

Precursors to the future

Data is generated 24/7 from multiple sources—intrusion detection and prevention systems, firewalls, routers, antivirus, hotspot, and more—in different formats, and in gigabytes or terabytes[1]. This has created large silos of security data. The resulting complexity makes it difficult to correlate, aggregate, and analyze information to recognize threat incidences. Understanding and analyzing un-unified data is also extremely time consuming.

Despite the presence of enormous security breach detection systems, the average breach detection gap, or dwell time, for most organizations is six to eight months, whereas the average data retention period of the security systems is two to six months[2]. Additionally, as data grows exponentially, the cost of solutions rises every year, as security systems are licensed by the byte. As a result, many chief information security officers spend more on their security budget, albeit without significant gains. Traditional tools are often not designed to scale horizontally, while the voluminous data generated needs to be ingested and retained with cost-effective scalable infrastructure.

Fixed rules configured in traditional security breach detection systems are too inflexible to counter the sophisticated techniques used by cybersecurity adversaries. As a result, thought leaders advise organizations to adopt statistical analysis and machine learning (ML) to enhance data security. However, with a variety of different technologies available, data scientists rarely standardize one technology or library for all problems. It is often difficult to integrate and utilize ML models with legacy security systems.

To address the complexity of data and the integration issues associated with security breach detection systems, organizations need a central platform for poly-structured security information[3]. The central platform must also offer features such as anomaly detection, contextualization, correlation, and search options combined with ML, to make the system efficient.

A new security umbrella

The concept of a specialized security data lake is relatively new and best described as kappa architecture[4].

At its core, a security data lake solution should comprise the following characteristics:

- **Data source agnostic:** The solution plugs in a wide variety of parsers, such as a grok or Java parser, to unify streaming log events into a standardized flat structure, as is the case in JavaScript object notation (JSON). This ensures that the ML correlation engine and downstream systems can relate or integrate the messages coming from different data points. These JSON files reside in Hadoop distributed file system (HDFS) and the search engines can perform a quick search on the indexed data present in HDFS.
- **Contextualization:** Further information can be contextualized to make them more relevant to the business. For example, an external IP address is enriched with GeoIP information (latitude/longitude coordinates, city, state, country, etc.) or Host details.
- **Intelligent behaviour-driven threat detection:** This is based on dynamic rules/probabilities and constant ML.
- **Centralized real-time search and UI dashboards:** These dashboards index all events and alerts and perform real-time searches.
- **Scalability:** Big data-enabled distributed platforms ingest and process large volumes of information at high speeds, as required.
- **Retention:** Offload PCAP, firewall, domain name system, windows events, audit log data from the security system, and SIEM to increase retention in a low-cost big data-based commodity infrastructure.
- **ML-based predictions:** These provide a huge data platform and the technical arm to understand historical context through ML and allows these models to be plugged directly into real-time pipelines.
- **Centralized view of threats:** Enterprises can view security threats, from organization-wide to individual risk activity, on a single platform.

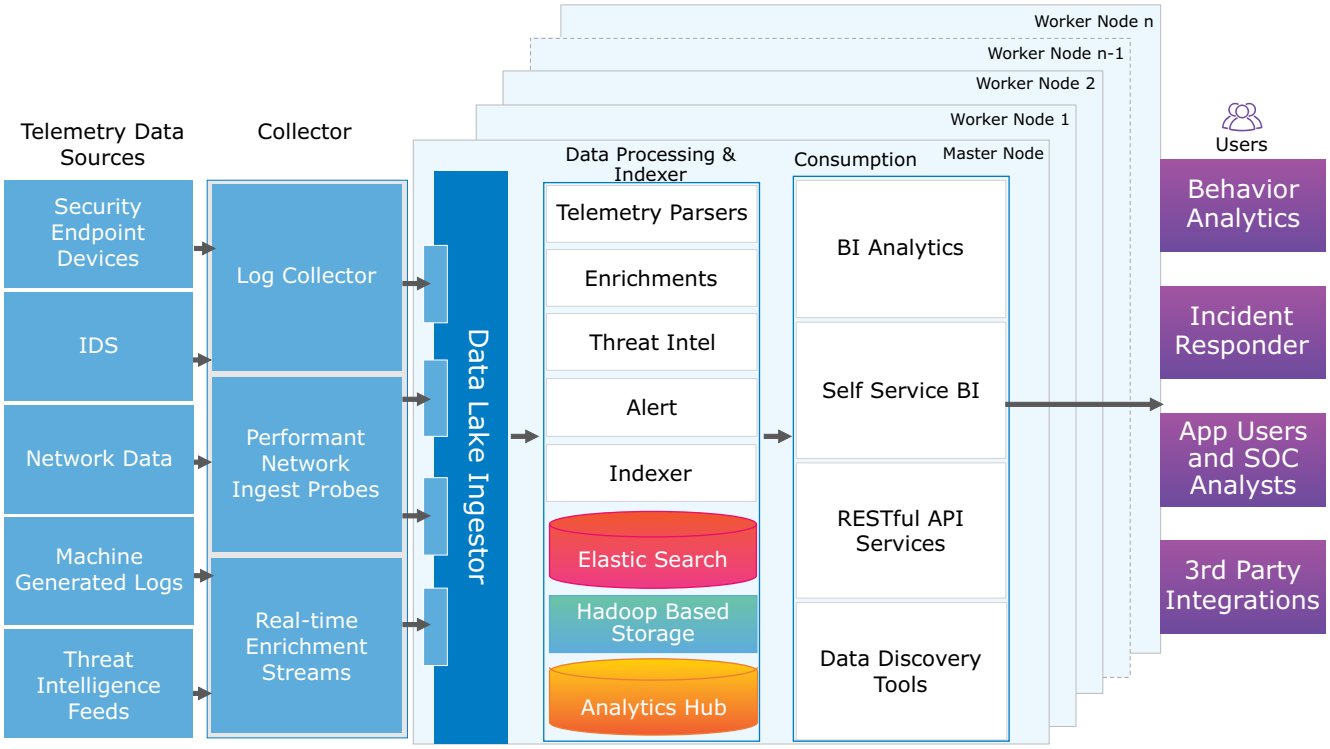

Figure 1: Big Data-Enabled Security Data Lake Architecture

A superior data processing solution

Raw events from telemetry data sources are streamed and published through the publish-subscribe component before parsing. Once the security data lake processing engine parses the raw data into a standardized flat structure, the correlation engine present downstream integrates the different data points.

The engine can parse only some fields of the log file as required by the business. Again, depending on the requirement, the security data lake can parse the telemetry data in three facets: collection-time parsing, batch parsing, and processing-time parsing.

Once the raw security telemetry event has been parsed and normalized, the next step is to enrich the different data elements of the normalized event, as better enrichment provides better context.

Analytical models can be run using the model-as-a-service (MaaS) pattern with the telemetry events flowing in, or telemetry events can be configured, which can initiate alerts based on different criteria. During this step, all enriched and labelled telemetry events are indexed and recorded in Hadoop for long-term storage. Such storage becomes a security data vault within the enterprise and the data can be examined with next-generation analytics.

There are different ways to access this data: full-text search, querying for slicing and dicing the data, querying data for real-time statistics and correlation, querying for deriving relationships between log records, or querying data for mining. Depending on data access requirements, organizations can also create a sub-storage system for enhanced data access.

Telemetry events stored in HDFS are indexed for elastic search-based configuration. Since data in the security data lake is normalized, enriched, and stored in standardized locations, a relational database layer can be built for business intelligence analytics. For real-time analytics, data can be accessed directly from HDFS or through a columnar database.

Security data lake provides the organization with a platform for real-time search. It also supports interactive dashboards to view alerts and correlate them to granular telemetry events that led to it. Big data-enabled security data lake provides organizations with a comprehensive view of enterprise security risk in a single pane of glass.

Conclusion

Though big data-based security data lake is a cost-effective and scalable solution, it is unlikely to completely replace traditional SIEM because of the latter's key capabilities such as a mature enrichment framework, an integrated ticketing and case management system, and inbuilt dashboards and reports designed especially for security. Organizations will benefit more if the SIEM is embedded in a data lake strategy. Data collected in the security data lake can be leveraged not only by security teams, but also by other teams within an organization that need to access the same data. Though big data-based security data lake can accurately identify cyberthreats, knowing what data to look for and how to store it are key points to be considered before implementing the solution.

Lastly, a security data lake solution is further beneficial to enterprises when it is offered along with big data[5], artificial intelligence, and blockchain[6]. Such a suite of solutions can act as silver bullets in the fight against cyberthreats. To illustrate, a big data-based data lake platform can train an AI by providing large volumes of data, while blockchain's immutability, transparency, and decentralization can prevent data from being tampering with[7].

References

[1] Gartner; Data-Related Issues Feature Among Top 2019 Risks for Internal Audit; November 15, 2018; https://www.gartner.com/smarterwithgartner/data-related-issues-feature-among-top-2019-risks-for-internal-audit/

[2] ComputerWeekly.com; Average attacker dwell time nearly six months for EMEA, study shows; April 4, 2018; https://www.computerweekly.com/news/252438158/Average-attacker-dwell-time-nearly-six-months-for-EMEA-study-shows

[3] Gartner; Why Your Security Data Lake Project Will FAIL!; April 11, 2017; https://blogs.gartner.com/anton-chuvakin/2017/04/11/why-your-security-data-lake-project-will-fail/

[4] Help Net Security; What is a security data lake?; January 30, 2018; https://www.helpnetsecurity.com/2018/01/30/security-data-lake/

[5] Info Security; Big Data and Cybersecurity - Making it Work in Practice; February 2, 2018; https://www.infosecurity-magazine.com/opinions/big-data-work-practice/

[6] Steel Kiwi; Using Blockchain Technology to Boost Cyber Security; https://steelkiwi.com/blog/using-blockchain-technology-to-boost-cybersecurity/

[7] Cloud Era Blog; Blockchain-driven Data Marketplaces: A Reference Architecture; May 31, 2018; https://blog.cloudera.com/blockchain-driven-data-marketplaces-reference-architecture/

About the Author

Sagarika Singh

Sagarika Singh is a solution architect in the Digital and Enterprise Transformation (DET) group of TCS' HiTech business unit. She works on next-generation data transformation engagements and has been part of several strategic solutions, consulting, and implementations for TCS' global clients. She holds a bachelor's degree in computer science and engineering from Utkal University, Odisha.

Contact

Visit the HiTech page on www.tcs.com

Email: HiTech.Marketing@tcs.com

Blog: HiTech Bytes

Subscribe to TCS White Papers

TCS.com RSS: http://www.tcs.com/rss_feeds/Pages/feed.aspx?f=w
Feedburner: http://feeds2.feedburner.com/tcswhitepapers

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is an IT services, consulting and business solutions organization that delivers real results to global business, ensuring a level of certainty no other firm can match. TCS offers a consulting-led, integrated portfolio of IT and IT-enabled, infrastructure, engineering and assurance services. This is delivered through its unique Global Network Delivery Model™, recognized as the benchmark of excellence in software development. A part of the Tata Group, India's largest industrial conglomerate, TCS has a global footprint and is listed on the National Stock Exchange and Bombay Stock Exchange in India.

For more information, visit us at www.tcs.com

TCS Design Services I M I 03 I 20