

Data-centric threat hunting

How AI/ML modeling
empowers a new generation
of threat hunters



Introduction

As threats become more sophisticated, so do the cybersecurity technology innovations that help us prevent, detect, and respond to them. While these tools and techniques provide valiant security analysts and threat hunters the chance for a winning advantage, they also create complexity and a “data dilemma” problem.

Enterprises generate, ingest, and store ever-increasing amounts of data from users, devices, sensors, and systems daily. Estimates range from gigabytes to terabytes per day, and by 2025, this amount is expected to reach 180 zettabytes.¹ This is inclusive of data generated by cybersecurity tools and solutions. Beyond just sheer volume, business data — and metadata — gathered from multiple disparate sources, in different forms and formats, provide varying depths of information, yet the efforts

to turn raw data into actionable, connected business intelligence is like moving mountains.

But the rewards are self-evident. Superior data utilization goes well beyond capturing transactional data and scanning for unwanted activity. By unlocking the power of data, an enterprise can examine its past, react more effectively to the present, and chart a course for the future that’s more secure and better protected.

¹ “Big Data Statistics in 2023: How Much Data Is in the World?” First Site Guide, 2023

AI/ML: A new foundation for data health

Using artificial intelligence and machine learning (AI/ML) tools, data analysts and scientists can find data nuggets and context that provide deeper insights to help security teams expose cyberthreats. While generic, out-of-the-box detections can help get less mature security organizations up and running, nothing beats using data directly sourced from your organization to provide significant contextual awareness.

Once data is normalized, the real cybersecurity work can commence, but looking for threat signals in all that data is like looking for a needle in a haystack. How can enterprises sidestep some of these challenges while reducing their risk and empowering highly effective threat hunting?

DataBee™, from Comcast Technology Solutions, enables data consumers and security teams to build improved AI/ML-based models to aid threat hunting. This guide will explore the details of using a data-centric approach for improved threat hunting using AI/ML modeling.

Building models using AI/ML to fight cybercrime is becoming within reach. As with any nascent innovation, there are multiple challenges to overcome:

-  Buying security products to mature your security program can quickly become expensive to do, and to maintain.
-  Cybersecurity experts with the right know-how are in short supply.
-  Data analysis is often complex, as each source may use unique syntax to represent and structure the data.
-  Data parsing is an intricate process, making any analysis or threat hunting challenging.
-  Normalizing the data is time-consuming but necessary to get its structure consistent for analysis.





AI/ML in threat hunting

Artificial intelligence and machine learning

Artificial intelligence and machine learning are terms that are found together regularly, and they're subject to a lot of misunderstanding and misuse. Let's start with a basic understanding of how these two terms relate:



Artificial intelligence (AI) refers to the processes and algorithms that simulate human intelligence, including mimicking cognitive functions such as perception, learning and problem-solving, visual perception, speech recognition, and decision-making.



Machine learning (ML) is a subset of artificial intelligence and is a way for a machine to imitate intelligent human behavior without explicitly being programmed. Machine learning is one way to use AI.

Why is AI/ML being talked about in virtually every industry – especially in the cyber community? It's the ability to thoughtfully cull through terabytes of data in a fraction of the time it would take a human or team of humans, coupled with increasing levels of comprehension and context with continued use, to make for a superior way to stay ahead of threat innovations.

Threat hunting

Threat hunting proactively seeks to identify adversary activity that existing detections or incident response programs miss. Large enterprises often have internal threat hunting teams, while smaller organizations may outsource the function to a third party due to cost.

Today, in most organizations, threat hunting is still a largely manual process. Threat hunters typically begin their investigations by developing a hypothesis following an actual or suspected security event or breach. Hunters then need to find evidence or artifacts of that incident using data pulled from multiple sources. These investigations can span days, weeks, or months. It can often be a laborious, tedious endeavor.

AI/ML accelerates — and elevates — threat hunting efforts

The more data and logs a threat hunter can review and correlate, the more successful they are likely to be. AI/ML can intelligently investigate terabytes of data more rapidly than humans, accelerating the process. More specifically, AI/ML helps expedite threat hunting by amplifying:



Speed and scale: Raw and processed data can be stored in a data lake, which is an ideal place to test and run AI/ML models. Modern cloud-based data lakes running on elastic compute engines can provide powerful performance and enable collaboration.



Data optimization: AI/ML can assist with optimizing data to ensure that there is no loss in data quality. This enables the data consumer (e.g., threat hunters) to trust its integrity when testing queries and conducting investigations. Data can be parsed and deduplicated at scale using AI/ML so it can be reviewed quickly to provide increased visibility and early detection of threats.



Context: Security events without business context require arduous event triaging and prioritization. AI/ML facilitates the evaluation of multiple baseline scenarios with deeper context, leading to better decision-making and fewer false positives, empowering threat hunters to maintain greater focus only on higher-priority threat signals.



Reduced cost: Security teams can use AI/ML to integrate and create cleaner datasets that reduce the computation strain on existing security analytics. By enhancing data upstream and leveraging data lakes, security information and event management (SIEM) tools can provide faster outputs without pushing the limits on storage capacity, ultimately reducing the pressure on CISO and CIO budgets.

Workflow for AI/ML modeling

Threat hunting requires a collaborative, yet programmatic, workflow approach. Data must be specially prepared and manipulated by data scientists and analysts in order for threat hunters to receive optimal value. The workflow stages include steps to:

- **Develop use cases.**
- **Capture and store data.**
- **Cleanse data.**
- **Transform data: parse, normalize, and enrich.**
- **Create baselines.**
- **Perform threat hunting.**

Develop use cases

Threats and threat actors are strategic, and your threat hunting program should be as well. Your security team should run tabletop exercises to identify relevant use cases. Threat hunters use AI/ML outputs from many sources: operating system logs, application logs, endpoint security logs, telemetry data, alerts, cloud logs, cloud platform and accounts information, user access logs, etc. They can use this data to pinpoint unusual or suspicious behavior and hypothesize use cases. These use cases may include specific or suspected threats not previously detected by in-house incident response teams or off-the-shelf security tools.

Capture and store data

To create effective AI/ML models for investigations, you need data. Data science teams strive to capture as much data as possible and store them for as long as possible to find trends and create relevant intelligence about the business. Multiple data sources are usually involved, including the organization's pertinent data sources, cloud-based storage services like AWS S3 or Azure Cloud Storage, and data lakes or warehouses.

Cleanse data

Once the data is captured, it needs to be cleaned. Data cleansing is the process of deleting or fixing improper, corrupt, or incorrectly formatted data. This is also the process whereby data is deduplicated and data fragments are addressed. For data teams, clean data is a top priority because the better the cleansing results are, the better and faster it is to extract insights from the data.

Transform data: Parse, normalize, and enrich

At this point, collected data from disparate sources is ready to be transformed into a standard format so it is more searchable and usable for data consumers. While there are many schemas to choose from, DataBee maps security data to the Open Cybersecurity Schema Framework (OCSF), an emerging vendor-agnostic schema that standardizes security taxonomy and aims to reduce the effort and time it takes to analyze data. During the transformation process, data will go through the following steps:



Data ingestion:

The collection of raw data from multiple sources into storage mediums



Data parsing:

The process of identifying patterns and extracting data from large datasets into another format



Data flattening:

The process of converting data, whether structured like firewall logs or unstructured Word docs, into fewer, more manageable datasets



Data normalization:

The process of turning data into a standardized, consistent schema. Examples of popular schemas are Elastic Common Schema and OCSF.



Data enrichment:

The process of enhancing data with relevant information and attributes to improve the analytical and operational value of the data

Create baselines

Once the data is transformed, data scientists collaborate with threat hunters to develop and define baselines. Typically, this means describing the environment or situation “right now” and searching for deviations to indicate malicious activity. Creating proper baselines requires expertise to know what attributes and data points to use and how to use them. For example, a baseline may have many attributes, but not all attributes may benefit the use case — or one data point might overlap. Therefore, security experience and expertise are vital to asking the right questions, interpreting the results, and creating the final baselines.

Perform threat hunting

With the baseline defined, threat hunting starts in earnest. The hunters launch queries against the data, looking for anomalies against the baseline. These queries might return hundreds or thousands of results that must be further parsed and analyzed. The next step is to validate the efficacy of the data and hunt. This involves the threat hunters trying to mimic the threat actors' behavior. And often, the testing implies that the baseline needs to be adjusted, so the whole cycle repeats.



Why engaging with data early in the pipeline is important

Engaging early with data sets the tone for the entire AI/ML interaction, improving data quality and model scalability.

Specific workflow steps like data cleansing and enrichment are long-pole activities, so allowing extra time can only be beneficial — and even necessary. Early involvement permits more time for data extraction and modeling. Thus, early engagement ensures ample focus by data scientists to optimize the data and for the threat hunters to engage with it.

Early engagement can also help set the parameters for budget and cost allocations. The cost of storing and analyzing massive amounts of data can be prohibitive and can escalate along with increased volume. Establishing a budget provides the guardrails and guidance to teams that help to control costs, and to prioritize projects and use cases.



Threat hunting use case example:

Lateral movement

For threat hunters, lateral movement refers to attackers exploring a compromised network to find vulnerabilities or assets of value. It is called lateral movement because once attackers gain access, they explore and pivot between systems undetected to achieve their objective, which may be espionage or a more sinister goal like installing remote access tools. An attack might begin with malware placed on an employee's device from a successful phishing email.

Detecting lateral movement without the help of AI/ML has traditionally been a time-consuming, repetitive process. First, the hunters would take inventory of their organization's attack surface, giving special attention to how a threat actor might target points of vulnerability. Hunters would then consider existing techniques observed in frameworks like MITRE ATT&CK that would help them write, test, and run various anomaly detection techniques to detect lateral movement.

With AI/ML, threat hunters can quickly automate and accelerate hunting techniques for lateral movement, even in massive, hectic, noisy environments.

Access to an integrated data layer consisting of large connected datasets enables threat hunters to cover attack vectors previously unconsidered. Without access to that data, there are visibility gaps where an attacker could hide. And this might prevent a threat hunter from a successful hunt altogether.

While the benefits of using AI/ML for threat hunting far outweigh manual detection techniques, it is not a panacea. It can assist with hunting and speed up the process, but it is only as good as the threat hunters' experience, models, and baselines. Hence, having a seasoned threat hunting team coupled with experienced data science and analytical teams is critical to success.

For example, hunters might look for suspicious lateral movement using Microsoft's Windows Management Instrumentation (WMI). Hunters will first identify the processes on all hosts with a parent process consistent with lateral movement. From there, they will use grouping and stacking to identify rare command line strings within the environment for further review and deep manual analysis.



AI/ML, threat hunting, and Comcast

Comcast is a Fortune 30 company that produces terabytes of data from diverse sources daily.

Comcast actively threat hunts against this ever-growing data landscape by leveraging a security, risk, and privacy data fabric that merges disparate data sources and feeds with organizational data and business. The result is proactive threat detection with near-limitless hunts, thus reducing the company's overall risk profile.

DataBee enhances threat hunting by enabling large organizations to conduct multiple, simultaneous, complex hunts from large-scale, historical datasets collected, normalized, and enriched into an authoritative source. DataBee removes compute constraints, enabling threat hunters to schedule intensive queries whenever necessary.




How DataBee can help organizations of any size with modern threat hunting

Up until now, AI/ML modeling for threat hunting has been cost-prohibitive. The initial cost and resource investment often exceeded the budget of many security organizations. On top of that, recruiting and training experienced data scientists and threat hunters adds another challenging complexity.

Comcast designed DataBee to overcome many of these financial and logistical challenges. DataBee is a cloud-native security, risk, and compliance data fabric to accelerate AI/ML initiatives. DataBee modernizes data management and provides integrated and enriched insights by connecting disparate data sources and feeds with organizational

data and business intelligence. With simplified and enhanced traceability for data ingestion, flexible enrichment options, and automated normalization, DataBee produces an enhanced dataset ready for high-performance analysis and reporting. Customers can stop choosing between capacity and security and ensure data quality by storing optimized data in a data lake.

DataBee benefits

-  Use data to drive cross-department collaboration.
-  Maximize data efficiency without sacrificing quality.
-  Gain back control and ownership of your data.
-  Stay ahead of threats and changing data privacy regulations.

DataBee was initially designed by the internal Comcast security, compliance, and data team. Through the implementation of a security data fabric at Comcast that spans across 150,000+ employees and over 1M endpoints, Comcast was able to maintain >50TB a day of data throughput and retain 10PB of security data for over a year while reducing \$10M in annual operational costs. Eliminating point products and optimizing SIEM ingestion with cleaner data enabled 65% higher fidelity alerts and 3X faster threat detection.

DataBee from Comcast Technology Solutions helps you get more out of your data

Comcast Technology Solutions (CTS), a division of one of the world's leading media and technology companies, is proud to offer DataBee™, our cloud-native security, risk, and compliance data fabric platform.

DataBee modernizes data management and provides integrated and enriched insights by connecting disparate data sources with business intelligence. Inspired by Comcast's internal security and compliance team, DataBee enables continuous compliance assurance and threat detection while optimizing data costs. At CTS, we invest in and test new ways to enable our partners to think big, go beyond, and lead the way in media, technology, and cybersecurity.

[MORE INFORMATION ON DATABEE →](#)

