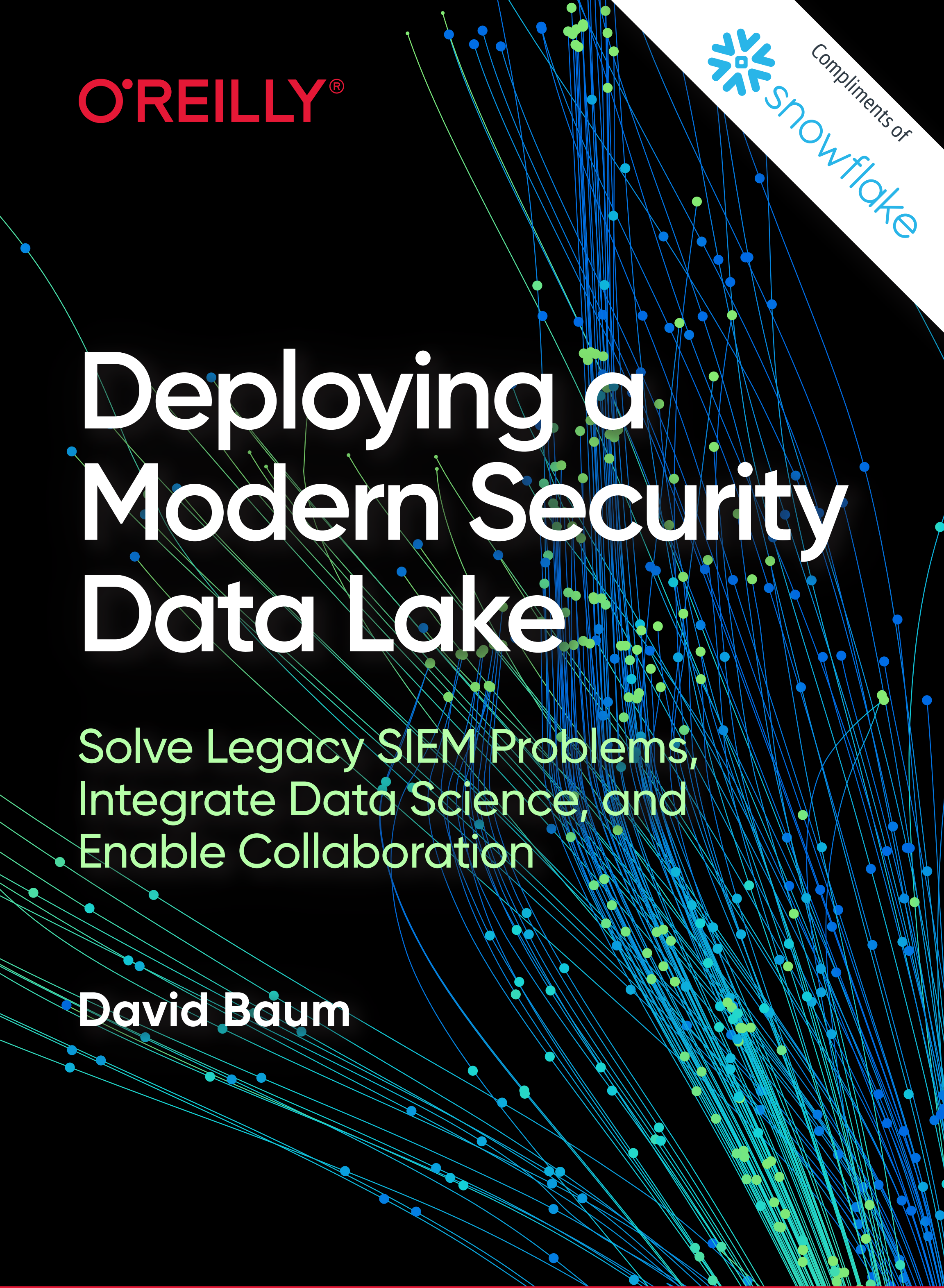
O'REILLY®



Deploying a Modern Security Data Lake

Solve Legacy SIEM Problems, Integrate Data Science, and Enable Collaboration

David Baum

REPORT

Deploying a Modern Security Data Lake

Solve Legacy SIEM Problems, Integrate Data Science, and Enable Collaboration

David Baum

Beijing · Boston · Farnham · Sebastopol · Tokyo

O'REILLY®

Deploying a Modern Security Data Lake

by David Baum

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Nicole Butterfield
Development Editor: Gary O'Brien
Production Editor: Kate Galloway
Copyeditor: nSight, Inc.
Proofreader: Jonathon Owen
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Kate Dullea

July 2022: First Edition

Revision History for the First Edition

2022-07-21: First Release

The views expressed in this work are those of the author and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

This work is part of a collaboration between O'Reilly and Snowflake. See our statement of editorial independence.

978-1-098-13495-2

[LSI]

Table of Contents

Acknowledgments

We would like to thank the following members of the Snowflake team for their valuable contributions to the report: Kelly Huang, Vincent Morello, Omer Singer, Haider Dost, Daniel Wyleczuk-Stern, Michele Freschi, and Martin V. Minner.

CHAPTER 1

The Rise of the Security Data Lake

Today's cybersecurity experts are overwhelmed. They are constantly on guard against malicious activity on their networks, from advanced malware infections to persistent threats, and from phishing schemes to SQL injection attacks. These external assaults are further complicated by the growing number of internal risks arising from simple errors, disgruntled employees, and outdated software configurations. Security experts must act on the assumption that all applications, services, identities, and networks are under threat.

For years, cybersecurity teams have relied on standalone security information event management (SIEM) systems that aggregate log data from firewalls, servers, network devices, and other sources. By pulling together these data points, analysts in the security operations center (SOC) can detect and respond to attacks as they happen, with the goal of mitigating threats quickly. Unfortunately, the process of identifying and investigating these incidents has failed to keep up with the complexities stemming from cloud computing, DevOps, work-from-home practices, and other emerging computing and lifestyle environments. Growing visibility gaps and insufficient automation prevent security teams from achieving their goals in threat detection and response, as well as tackling other important use cases such as regulatory compliance and vulnerability management.

In addition, as data from these activity logs grows in complexity and scope, most organizations find they can analyze only a small fraction of these vast repositories. Legacy security tools have limited data management capabilities and restrictive data storage allocations, which hinder the effectiveness of forensic investigations. Meanwhile, because of increasingly complex regulatory requirements, security professionals must help their organizations comply with strict data privacy regulations governing the creation, storage, and use of consumer data. This adds to the already onerous task of monitoring corporate information systems to avoid unauthorized incursions—before data is lost, trust is breached, or customers become aware of performance issues.

Security teams need a faster, easier way to get the data they need so they can stop bad actors before attacks escalate into breaches. Maintaining strong edge controls, such as endpoint detection and response (EDR) software and secure access service edge (SASE) systems, is an important part of network security, but savvy attackers know how to penetrate these virtual perimeters. To minimize risk, security operations teams must modernize their cybersecurity systems so they can detect, analyze, and even predict potential threats quickly and effectively to thwart breaches when intrusions do occur.

A *security data lake* is a specialized data lake designed for collecting and manipulating security data. This report describes how the security data lake model can complement or replace the traditional SIEM model. It also describes how to create a modern security data lake with an organization's existing cloud data platform to deliver comprehensive visibility and powerful automation across multiple security use cases.

Understanding the Limitations of the Traditional SIEM Model

SIEMs monitor and analyze data from users, software applications, hardware assets, cloud environments, and network devices. These information systems allow security professionals to recognize potential security threats and vulnerabilities before they have a chance to disrupt business operations. The data is collected, stored, and analyzed in real time, allowing security teams to automatically monitor logins, data downloads, and other activities, as well as track and log data for compliance or auditing purposes. Security experts write and maintain rules to manage alerts from servers, firewalls, antivirus applications, and many types of equipment sensors. This event data from machine logs and other network sources is stored in a database and presented via monitoring dashboards.

The traditional SIEM model was adequate when corporate data and applications resided in on-premises data centers. Today, as organizations migrate information systems to the cloud, adopt software as a service (SaaS) applications, and deploy an immense variety of web and mobile applications, they generate much more data. Every transaction—often every click, swipe, or tap—generates a record, leading to a deluge of log data. SIEM systems bog down under the burgeoning load of data arising from these cloud-driven log sources as well as from containerized systems such as Kubernetes.

Popular SIEM solutions are designed to search activity logs, overlooking complementary data sets such as asset inventory and configuration records. To supplement security logs, security analysts may also gather *contextual information* about the employees of the organization they work to protect, such as user location, device type, and job role. However, SIEM solutions can't ingest these other enterprise data sources, so they lack the ability to use complementary and contextual data sources to automatically weed out false positives (noisy alerts) from false negatives (undetected threats). The SIEM can only ingest log data and only in limited quantities. Furthermore, traditional SIEM providers offer only rudimentary analytics via their proprietary search languages.

SIEM solutions are also prohibitively expensive to use with high-volume data sets, such as cloud activity logs and endpoint forensic data. As a result, potentially important security data is kept siloed in low-cost, archival storage media, commonly known as *cold* archives because the data is accessed infrequently. By contrast, *hot* data resides in a readily accessible state for instant query and analysis, but generally only in limited quantities.

These restrictions stem from a basic problem: most SIEM solutions emerged when corporate applications and data resided in on-premises data centers. Most of these SIEM offerings now operate in the cloud, but they lack a true cloud multitenant architecture. These "cloud-washed" solutions can't take advantage of the near-unlimited storage and computing power the cloud offers. SIEM searches can take minutes or even hours to complete, and it is difficult to scale these deployments to match the steady growth in log data that arises from today's mobile and cloud-based information systems. SIEM solutions are also expensive because of high software license costs and excessive data storage costs. In turn, these solutions are constrained by limited retention windows and can't combine structured, semistructured, and unstructured data into a single repository. Figure 1-1 sums up these limitations.

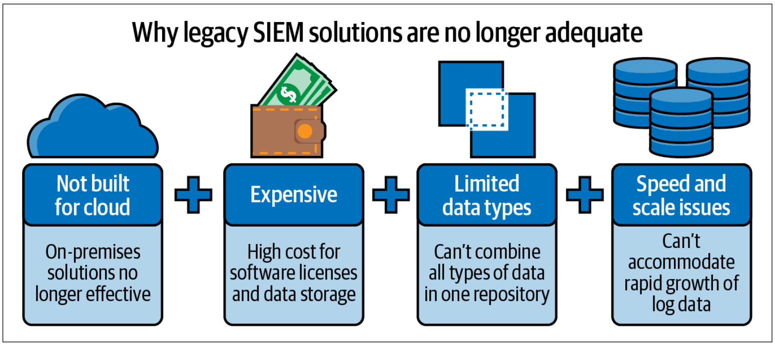

Why legacy SIEM solutions are no longer adequate
Not built for cloud
On-premises solutions no longer effective
Expensive
High cost for software licenses and data storage
Limited data types
Can't combine all types of data in one repository
Speed and scale issues
Can't accommodate rapid growth of log data

Figure 1-1. In the era of mobile computing and cloud-based applications, yesterday's SIEM tools are showing their limitations

Expanding Your Analytic Horizons

Every time an alert fires or a breach investigation is launched, security analysts must quickly identify the attacker's entry point, establish the blast radius, and validate whether or not an incident has been properly mitigated. However, it's hard with traditional SIEM solutions to synthesize insights from activity logs with contextual data sources, restricting the security organization's ability to effectively detect and respond to genuine threats. With traditional SIEM, monitoring activities remain largely manual, and the detection/response cycle often isn't quick enough to adequately thwart determined attackers.

According to the 2022 Verizon Data Breach Investigations Report, the time from an attacker's first action in an event chain to the initial compromise of an asset is typically measured in minutes, whereas the time to discovery is often measured in days, weeks, months, or even years.

Many reasons for this poor response time exist, beginning with the obvious: there are simply too many systems, devices, and applications that generate contextual data in disparate places. When these data sets are siloed, SIEM solutions are prevented from detecting a threat in one or more information systems.

Forward-looking security teams see the value in storing contextual data from dozens of sources in a single repository so they don't have to turn to many different places to find the right information. Consolidating security data streamlines investigations by allowing security teams to use robust analytics and data science methods to spot suspicious activity and respond to threats. A centralized repository makes it easier to apply modern business intelligence and predictive analytics capabilities to the data, dramatically improving the search-only interfaces that characterize traditional SIEM systems. This approach helps security teams overcome the negative impacts of SIEM solutions.

Reviewing Security Data Lake Prototypes

Most security teams rely on their SIEM solutions as a centralized data platform, pulling in data from various "point" products—specialized software solutions that address specific security use cases—to support threat detection and investigations. However, as more organizations move their information systems to the cloud and adopt a wide variety of SaaS applications, traditional SIEM solutions can't keep up with the complexity and volume of data storage. Because it is cost prohibitive to store large amounts of data in traditional SIEM products, organizations must decide which limited data they can collect from security sources and how long they can keep it available in an active, searchable, hot database.

As early security data lake prototypes gained prominence, cybersecurity teams saw the potential to leverage robust data platforms to power their immense security workloads. Early security data lakes were built using Hadoop. However, these open source systems required complex development and maintenance to get them up and running. Hadoop infrastructure also required specialized experts to implement, manage, and scale—skills most security teams don't have.

Other organizations cobbled together security data by storing some of it in a cold data archive, such as an Amazon Web Services S3 cloud storage bucket. This approach solved the data storage problem, but because no direct integrations among these disparate environments existed, these security data lakes required lots of tinkering. For example, organizations had to constantly restore data from a cold archive into an active or hot database to conduct analytics.

Introducing the Modern Cloud Security Data Lake

Security professionals who attempted to supplement their SIEM solutions by building a security data lake from scratch often found themselves mired in expensive projects with limited results. They learned that a complete security data lake involves much more than loading archival security logs and applying general-purpose analytics. Organizations must also overcome unique challenges related to ingesting, enriching, and formatting the data of many types from many sources for specific security use cases, including assessing the most pressing threat detection requirements and addressing those requirements with custom queries, machine learning models, and data visualizations.

To understand the power and potential of a modern security data lake, let's review the attributes of a general-purpose data lake—a versatile repository designed to store large amounts of data in native formats. This data can be structured, semistructured, or unstructured, and it can include tables, text files, system logs, and other sources. To maximize flexibility, data lakes do not impose a schema on the data when it is captured. Instead, the schema is applied when the data is extracted from the data lake, allowing for multiple use cases on the same data.

Just as a general-purpose data lake allows analysts and data scientists to consolidate many types of data for analysis and visualization, a modern cloud security data lake enables security teams to use one system to analyze many kinds of data. A security data lake should be able to store and manage data from single-, multi-, and cross-cloud environments that encompass an immense variety of SaaS apps along with data entrusted to public cloud providers such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP).

Every security product, network device, and computer on a network creates its own logs. Instead of forcing security analysts to manually gather and separately analyze data from all these siloed systems and devices, a modern security data lake includes data pipelines that pull it all together to enable consolidated analytics. Centralizing these logs in a security data lake simplifies threat investigations and other cyber use cases such as control validation, identity and access, and vulnerability management.

Harnessing the Power of a Cloud Data Platform and Connected Ecosystem

Early security data lake implementations resulted in a swamp of data that security analysts could not readily leverage for investigations. Modern security data lakes are enabled by a cloud data platform that can scale up and down automatically based on fluctuating workloads. As shown in Figure 1-2, these platforms provide inexpensive storage for structured, semistructured, and unstructured data, which is important for security use cases, and they uphold strong control and management capabilities to govern how users access the data. They also offer near-unlimited compute power and a growing ecosystem of connected applications, providing off-the-shelf capabilities complete with API integrations, purpose-built interfaces, and near-immediate access to up-to-date security content via data marketplaces.

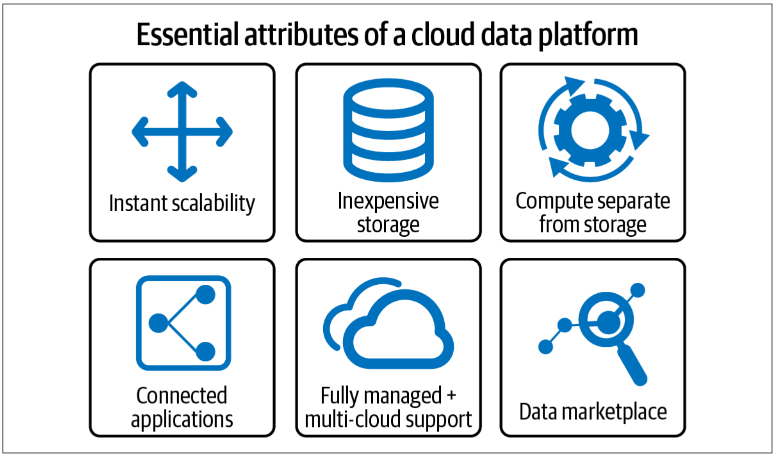

Essential attributes of a cloud data platform
Instant scalability
Inexpensive storage
Compute separate from storage
Connected applications
Fully managed + multi-cloud support
Data marketplace

Figure 1-2. A modern cloud data platform provides unique capabilities for creating security data lakes

With virtually unlimited cloud data storage capacity, security teams are no longer constrained by the data-ingestion and data-retention limits imposed by traditional SIEMs. They can store all their data in a single platform and maintain it all in a hot, readily accessible state. Threat intelligence data collected from SIEM logs and other security sources can be joined with other data sets to reveal the full scope of an incident from historical records.

As you will see in the chapters that follow, if you build your security data lake with a modern cloud data platform, you will obtain a data management solution that extends far beyond typical SIEM use cases to support many other parts of an advanced cybersecurity strategy. Your security team will be working within the same data platform as your other data teams and gain instant access to contextual data sets without inflicting more overhead. Your security team can partner with data professionals throughout the enterprise, with everybody working within a standards-based environment (in contrast to other security solutions that use proprietary languages and formats). This freedom allows your security operations center to apply all types of business intelligence and data science tools to the security discipline.

NOTE

Security data lakes apply advanced analytics, near-limitless cloud storage, and near-infinite elastic computing to the task of security analytics. They allow security teams to access a universal data repository that combines security data from across the enterprise into one system, making it easier to evaluate alerts and understand attack details. Powerful analytics help these security professionals detect and respond to threats, while security content and visualizations in connected applications help them to be more effective.

Summary

Attack surfaces are expanding as enterprises increasingly rely on complex, multi-cloud environments. Unfortunately, legacy SIEM solutions fail to enable effective threat detection and response in these diverse IT settings. These outdated solutions are plagued with data storage and retention limitations, along with poor scalability and slow query performance. As a result, many security teams can't easily determine what is happening across their organization's infrastructure. These limitations also limit historical reporting and data science initiatives because each set of security logs ingested into the SIEM is available for a limited period, typically 90 days or less. Effectively securing your environment is difficult when you have access to only some of your security data, some of the time.

To overcome these limitations, a growing number of organizations are moving their security data and SIEM workloads into security data lakes. A well-architected security data lake, based on a scalable cloud data platform, eliminates data ingestion and retention limits. A modern security data lake powers robust analytic capabilities on top of the cost-efficient data storage capability, which greatly reduces data management overhead and the manual investigation processes that traditional SIEM platforms require. Finally, aligning the security organization with the rest of the enterprise provides an opportunity to protect the business while enabling growth and innovation.

CHAPTER 2

Implementing a Security Data Lake

As explained in Chapter 1, today's cloud data platforms can power security data lakes that help you automate and greatly expand many cybersecurity tasks. But how do you gracefully transition from yesterday's SIEM-centric environments to embrace today's modern alternatives? You can't simply "lift and shift" your old data center security methods, because they no longer address the scale and complexity of today's multifaceted threat landscape. This chapter describes how to implement a security data lake to efficiently gather all your data, expand visibility into security risks and incidents, and automate responses to mitigate threats. This process has three primary phases:

1. Assess your current state.
2. Collect and migrate data.
3. Establish and verify analytics.

Phase 1: Assess Your Current State

During the assessment phase, you must answer several key questions:

- Which threats present the greatest risk to your organization?
- What are the key solutions being used to mitigate those risks?

- How is data from these sources being used today?
- What are the biggest challenges and gaps to success with these use cases?

For example, a review of an organization's threat models may point to source code theft as the top risk. Attacks targeting developer laptops may be considered the most likely vector for this threat. The relevant existing solutions might be the endpoint protection agents on developer laptops, and controls on the managed source code repository. However, data from those solutions may be siloed in the respective sources and not used together for identifying and spotting attacks against developers and their code. Based on this realization, security professionals might decide the best way to mitigate this risk is to apply behavior analytics and to support incident response measures across a combination of endpoint and source code repository activity logs. This assessment would inform the scope of the initial security data lake implementation.

As you assess the current state, don't limit the conversation to identifying threats and managing risk. Use this opportunity to talk to various members of your user community about the difficulties and limitations of their current tools and processes. Ask them which threat detections provide essential coverage for your organization. Have them delineate the manual activities they would most like to automate. Here are some additional questions to ask at this stage:

- What are the key security logs and data sources you depend on?
- How do you leverage this data?
- What data sources are you not collecting currently?
- What data sets are archived and not readily accessible?
- How would centralizing these data sources be helpful?
- In what ways can you collect data from these sources?
- Does your current SIEM solution provide the investigative capabilities you need? If not, in what ways is it lacking?
- How many detection rules are active across the environment? How many of these are prebuilt versus custom-developed?
- Can the team develop detections in house? Does it have adequate skills for data modeling and data engineering?

If you have your eye on a particular cloud data platform, research which cybersecurity vendors are part of that platform's ecosystem. Is there an ecosystem partner that offers the data sources you need? Does it offer the flexibility you'll need for data ingestion, threat detection, and advanced analytics? Will you depend on the vendor to supply new content in response to new and evolving attacker techniques? If so, investigate the skills and capabilities of that vendor's research team.

Remember, the ultimate goal of any security data lake project is to establish a single source of truth for your security program. At the same time, each data source and use case has its own nuances. Collecting network firewall logs is a different process than bringing in permission profiles from an identity management system or asset inventory records from a change management database. Does your chosen security data lake ecosystem accommodate all your essential data sources and priority use cases?

Phase 2: Collect and Migrate Data

Once you have determined which internal and third-party data sources to include in your security data lake, you must design and implement a data collection architecture to ingest that data. Consider using a design document to list the various categories of data sources such as SaaS applications, cloud infrastructure services, and security products, along with the integration methods you will use to migrate data from all these sources. Common integration technologies include application programming interfaces (APIs), direct data sharing, general-purpose data pipelines, and security-specific data collection solutions. Where possible, leverage prebuilt integrations from vendors and partners to ingest your data sources in order to minimize the connectors you develop and maintain. Once again, take a close look at the security data lake platform provider and its key partners. Can they automate these important tasks for you?

Data sources that are commonly streamed into the security data lake come from endpoint detection and response (EDR) solutions, firewalls, servers, email gateways, cloud infrastructure, and network flow taps. The diversity of these sources and the potential value of analyzing them together has led to the rise of *extended detection and response* (*XDR*) solutions, many of which are designed to work with a customer-owned security data lake.

After you have taken inventory of these common data sources, identify additional third-party data sets such as threat intelligence feeds available within the cloud data platform's ecosystem. Some data platforms have established marketplaces in which it is easy to identify available solutions for building and populating a security data lake. Identify your most critical log sources and migrate each of them incrementally, according to your priorities. Start with a small proof of concept and expand as your organization gains experience with its security data lake and the new security methods it enables.

Phase 3: Establish and Verify Analytics

Unlike traditional SIEM solutions, security data lakes provide schema-on-read capabilities that reduce the upfront burden of modeling and parsing data. As with a standard data lake strategy, data can be initially loaded in its raw state (see Figure 2-1).

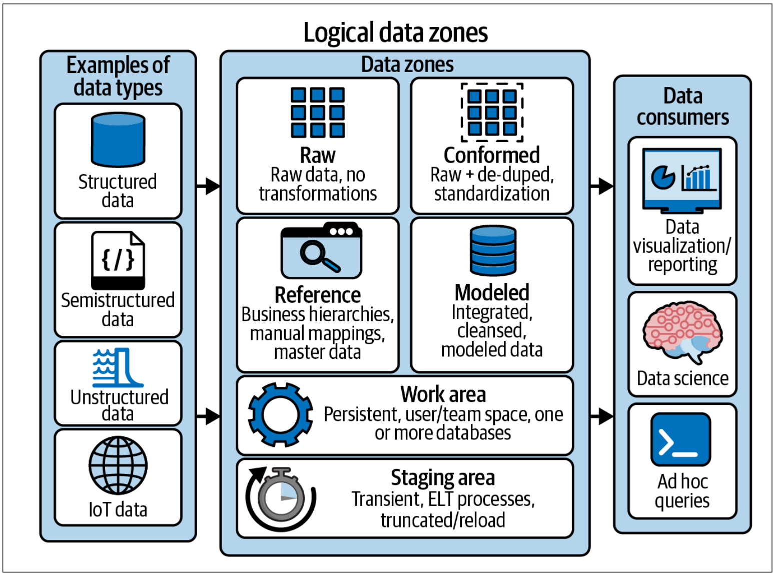

Logical data zones
Examples of data types
Structured data
Semistructured data
Unstructured data
IoT data
Data zones
Raw
Raw data, no transformations
Conformed
Raw + de-duped, standardization
Reference
Business hierarchies, manual mappings, master data
Modeled
Integrated, cleansed, modeled data
Work area
Persistent, user/team space, one or more databases
Staging area
Transient, ELT processes, truncated/reload
Data consumers
Data visualization/ reporting
Data science
Ad hoc queries

Figure 2-1. Data lakes are often designed to move data through various zones as the data is prepared for use. The same approach can support threat detection, hunting, and incident response.

During the streaming phase or after loading, automated transformations are applied to normalize, enrich, and clean the data. Raw events can still be queried directly as long as they are in valid JSON or XML formats.

Once the in-scope data is being loaded, be sure to test it for completeness and usability. Can you see the devices and networks you expected to see? Are commonly used fields nested deep in raw JSON, or are they readily accessible to analysts? Run through the kinds of triage and investigations your team members rely on before you consider the data collection is complete. Then, test your detections to make sure they work properly in the new environment. During this phase, you should attend to the following:

- Assessing and maintaining data quality
- Enabling and testing out-of-the-box alerts and queries from ecosystem partners
- Migrating custom detections from legacy systems
- Optimizing the schedule and frequency for threat detection rules
- Measuring and reporting metrics for security operation effectiveness using business intelligence (BI) tools

As you develop and test SQL queries that cover your end user's expected questions, make sure to consider the less frequent but potentially challenging queries and edge cases. You should also monitor query performance at production scale to make sure it is adequate for each use case. As you work with the users you identified in Phase 1, provide them with sample queries, and make sure those queries return the results your users anticipated.

Validate your preparedness with penetration tests that simulate attacks relevant to the threat models you have identified during this iteration of the security data lake buildout. For example, you might want to test for basic malware downloads, API keys that are compromised, and lateral movement of an attacker through particular domains. The goal is to identify gaps and issues in visibility, detection logic, and IR search queries. One of the advantages of a security data lake is its inherent support for engineering processes such as testing, measurement, and "detection as code," systematically reducing risk for the organization.

Finally, consider how you can further empower your users and other stakeholders with actionable dashboards and metrics. In a production security data lake, self-service BI dashboards meet most query requirements for business users that contribute to the security posture of the enterprise. As you shift from spreadsheets and slide decks to an enterprise BI tool, you'll find the audience for security dashboards extends beyond security analysts to include governance, risk management, and compliance (GRC) auditors, HR professionals, and even members of the C-suite.

Best Practices for a Successful Implementation

Pay attention to these guidelines to direct the implementation process:

- Identify the top security issues you hope to resolve.
- Don't try to collect all your data at once.
- Select the right security data lake vendor and partners.

Roles and Responsibilities

Many types of technology professionals play a role in cybersecurity initiatives.

Chief information security officers (CISOs), chief information officers (CIOs), and vice presidents of cybersecurity must keep their eye on the high-level risks and challenges. They are motivated by the following factors:

Reducing overhead
: Implementing a security data lake greatly simplifies the security program architecture by eliminating cold tiers of data. Many security leaders see the value in minimizing data silos across different tools, such as unifying log data spread across SIEM storage and self-managed cloud buckets.

Responding rapidly to threats
: When all security data is securely centralized alongside business data sets, security teams can streamline incident response playbooks and mitigate threats faster.

Security analysts, security engineers, and security architects strive to achieve cybersecurity best practices, with attention to these key factors:

Improving alert fidelity

Security operations center (SOC) teams strive to minimize both false positives and false negatives. This effort is supported by using SQL to create detections that span multiple data sets as needed for contextualization within the alert logic.

Accelerating investigations

Better tools for data ingestion, integration, and advanced analytics eliminate manual investigative work by gathering relevant details, evidence, and intelligence on behalf of the analysts.

Database professionals—from database administrators (DBAs) to chief data officers (CDOs)—work together to address these critical tasks:

Eliminating data silos with existing solutions

When security data is siloed from the rest of the organization's data, it makes it harder for data teams to play a meaningful role in protecting the business. As a result, many cybersecurity teams are stuck with rudimentary analytics and canned dashboards that don't give them the insight they need. A modern cloud security data lake uses standard analytics tools and languages familiar to data professionals. A partnership between data and cyber teams means existing investments in data pipelines and data platforms can be extended to the security program.

Applying advanced analytics to security use cases

Data teams have been driving substantial productivity gains and innovation for departments such as marketing and finance. This progress has been held back in cybersecurity by legacy solutions that only support niche, proprietary search query languages. Collaboration between data scientists and security domain experts within a modern cloud data platform unlocks opportunities in threat hunting, anomaly detection, and risk forecasting that can transform the security posture of an organization.

Summary

Modern cloud data platforms enable security data lake initiatives that require less effort and are faster to implement than traditional security solutions. However, successful implementations require careful planning and benefit from the following best practices. To implement a security data lake that makes effective use of the enormous volume and complexity of your security data, remember these key steps:

- Take stock of your most pressing needs and match them against your current security stack to identify gaps in visibility or capabilities.
- Unify security data sources and enterprise data using built-in ingestion utilities within the cloud data platform or via an ecosystem partner's prebuilt connectors.
- Create a data model for accelerating analytics on security data, and use schema-on-read to query collected data in its raw state.
- Partner with the data team within your business to utilize existing data solutions and collaborate on a data-driven security strategy, including business intelligence (BI) reporting and behavior analytics.
- Investigate incidents, hunt threats, and proactively reduce risk by querying the security data lake both directly and through purpose-built interfaces in connected applications.

CHAPTER 3

Connecting Best-of-Breed Security Applications

As explained in Chapters 1 and 2, consolidating data in a modern security data lake eliminates data silos and accelerates time to value for all types of cybersecurity initiatives, from identification and protection to detection, response, and recovery. These cloud-built solutions allow security teams to apply powerful analytics to log data and other security-relevant information, all maintained as a single source of truth. This consolidated, single source of truth improves visibility into all relevant data, leading to higher-fidelity insights and better security outcomes.

However, while cloud data platforms support cost-effective analytics at a massive scale, they don't include all the security integrations, interfaces, and content that security teams need. To complete the stack, cloud data platform providers work with third-party software vendors that specialize in solving security use cases. These security applications typically include out-of-the-box connectors, purpose-built interfaces, and detections that are frequently updated as the landscape evolves. Your chosen cloud data platform should easily integrate with these connected applications so you can quickly add their capabilities to your security data lake. These solutions will empower your security team to transition from the siloed solutions of the past as you perform deep investigations and quickly resolve security incidents.

Today's purpose-built solutions, whether they are labeled as open SIEMs, XDRs, or security analytics platforms, handle everything from data ingestion to incident response, complementing or replacing the more limited scalability and analytic capabilities of traditional SIEMs. The goal of these connected ecosystems is to simplify the creation, operation, and maintenance of your security data lake.

Connected security apps typically include a high-level GUI interface for security analysts who aren't yet comfortable with SQL. Tailored interfaces can be helpful for advanced users as well when they support graph-type navigation or point-and-click pivots. Security teams can accelerate their implementations by leveraging out-of-the-box integrations, content, and visualizations. They can also consolidate all their enterprise and security data into a single location and take advantage of advanced analytics for detection and response (see Figure 3-1).

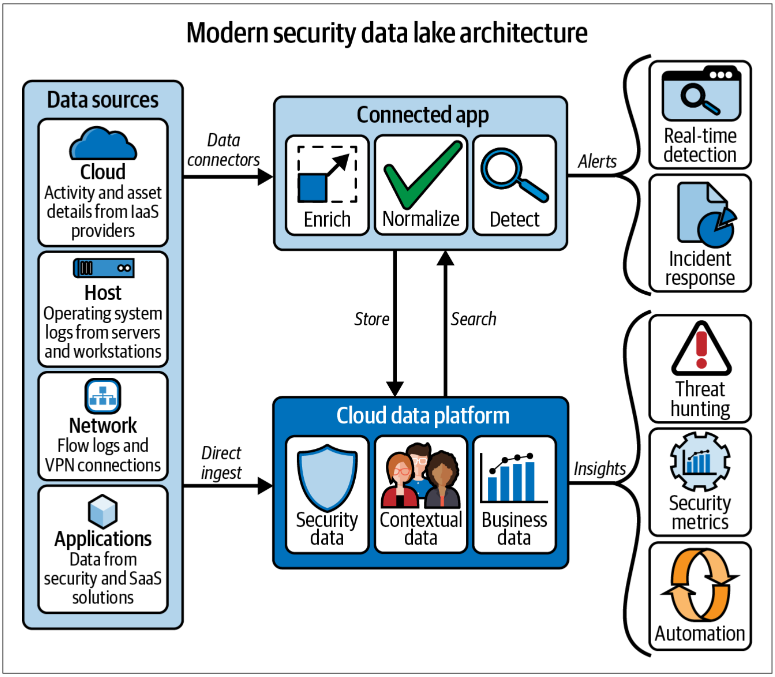


Figure 3-1. A modern cloud security data lake supports near-unlimited ingestion and retention of data in a single tier for all sources and can accommodate multiple security use cases with connected applications

Understanding the Connected Applications Model

Most SaaS applications maintain customer data in a traditional *managed services* model, in which the SaaS vendor is responsible for the application code as well as the customer data. To access that data, customers must connect to the SaaS vendor's APIs, placing the burden on the customer to build and maintain data pipelines that provide visibility into the applications they use. Creating and maintaining these pipelines can quickly become an onerous task because many organizations rely on dozens or hundreds of SaaS applications. Security teams must manage multiple pipelines and long integration backlogs to obtain the data they need. Furthermore, each SaaS application creates its own data silo, leading to the problems of data fragmentation described in Chapters 1 and 2.

The connected application model hands control to the customer. Modern SaaS vendors that utilize this model build their SaaS apps on a cloud data platform. Customers that build their security data lakes *on that same cloud data platform* can deploy their SaaS applications directly on their own data platform environment and maintain control of that data without the vendor having a separate copy. This approach is called the *connected applications model* because of its natural division of responsibility: the SaaS vendor creates and maintains the software code while each customer can maintain its own data in its own security data lake.

In the connected applications model, SaaS vendors handle the technical nuances of integrating their data with each customer's security data lake. The vendors establish data pipelines to the cloud data platform, through which they can connect to individual customer instances. Each customer maintains its own data and enforces its own security and governance controls. Custom API connections are not necessary.

In addition to acquiring standard, out-of-the-box security capabilities from connected applications, customers can use rules built in their SIEM systems, as well as develop and run custom rules, reports, and analytics using SQL, Python, and other popular languages.

NOTE

Legacy SIEM solutions restrict security teams with rudimentary dashboards, but security data lake vendors offer a broad ecosystem of connected applications to complement and extend their core capabilities. Customers can draw on the resources of these ecosystems to leverage out-of-the-box API integrations, prebuilt user interfaces, and standards-based security content, detections, automated playbooks, and more. These integrated tools also simplify tracking and reporting security metrics.

Context Matters

Most security solutions offer users some level of analytics and reporting. The problem facing the users when they log into each point solution is that they can only see endpoint activity or vulnerability findings in a silo. This makes it hard, if not impossible, to achieve high-fidelity insights or establish automated workflows. Security professionals must learn the unique nuances of each environment before taking action within the data environment as a whole.

A security data lake that combines data from all logs, users, assets, and configurations into a cohesive repository makes it much easier to understand the contextual relations among data elements, greatly expanding the possibilities for automation. For example, a detection engine can connect the dots to uncover threats that would have been impossible to spot by looking at individual sources on their own. A real-world example would be a compliance analyst joining employee termination records from the HR application, permission policies from the identity directory, and authentication events from sensitive systems to identify users who accessed sensitive data or resources, and to determine whether they abused that access. This versatile model opens new opportunities for threat hunting as well as other security tasks that benefit from a holistic view.

This architecture empowers not just analysts but also a new generation of applications that intelligently connect information across data sets, merging security data with other enterprise sources. Autonomous threat hunting, permission rightsizing, and security control validation are some examples of connected application use cases that can take advantage of the context established in your modern security data lake.

Counting the Cost of Connected Applications

When evaluating cloud data platforms and how they work with connected applications, pay close attention to the pricing models that vendors offer so you can select the ones that best fit your requirements. Common pricing plans include ingestion-based models, subscription models, and consumption-based models.

Ingestion-based pricing is typically based on how many gigabytes of data you ingest each day, up to a given threshold. You pay a flat rate for your respective tier of usage, even if you don't ingest the allotted amount of data. This is the most common model in the SIEM space, and it can often drive up costs prohibitively when you operate these solutions at full scale. However, a modern security data lake can benefit from the architecture of its underlying data platform and offer ingestion-pricing at levels low enough to store a multitude of data types and volumes for full visibility.

Subscription pricing offers a fixed, flat-rate price for a specific period, such as per month or per year. A tiered structure allows you to subscribe to various levels of features at various price points. You pay for these products and services on a per-seat basis for a set period of time. However, you must estimate upfront how many seats or licenses you will need, often before you fully understand how your organization will adopt and use the solution. This is the most common model in cloud security posture management (CSPM) and vulnerability management. The security data lake portion of the connected application has a separate cost that you should factor into total cost calculations.

With a *consumption (usage-based) model*, you pay according to the compute and storage resources used, for example per second or per byte. This pricing model includes the cost of storage (such as per terabyte per month) and the cost of compute resources used to run queries, ingest data, or perform other data processing services (billed according to actual usage). This model can dramatically lower costs because you don't pay for unused capacity.

The consumption model reduces overall costs by aligning spending with actual activity. Cybersecurity work has slow periods interspersed with periods of intense urgency. For example, when a suspected breach occurs, SOC leaders demand fast answers, which justifies a temporary spike in spending. A security data lake allows you to accelerate investigative queries by six times or more by instantly scaling cloud compute resources. Some cloud data platforms allow you to provision distinct compute clusters to each team or workload so you can avoid degrading performance during high concurrency and more easily attribute costs among designated domains, departments, and cost centers.

Controlling and Securing Data from Connected Applications

Building your security data lake on a cloud data platform allows you to take control of your data. By controlling the underlying data platform, your security team can scale up compute resources elastically to enable powerful analytics during threat hunting or incident response. The process of deploying a connected application usually involves creating a dedicated user for the app within the cloud data platform account, then configuring that user's credentials into the connected application so that it can interact with the data platform. The connected application can use a continuous ingestion service to load data into the security data lake, as shown in Figure 3-2.

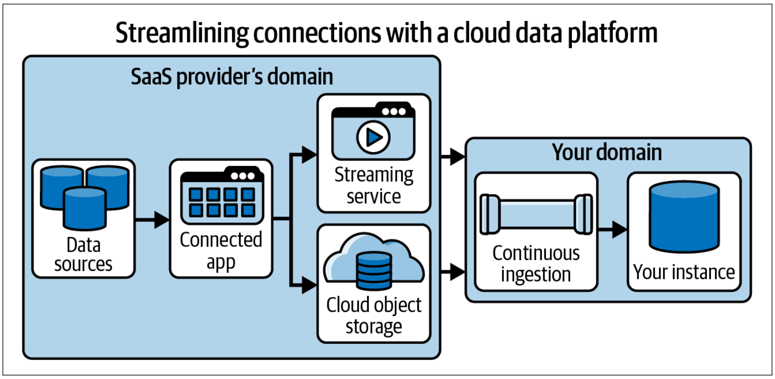


Figure 3-2. A cloud data platform serves as a natural convergence point for connected applications, making it much easier to acquire and integrate data from many SaaS providers

Enhancing Data Security

A unified source of truth can be shared by multiple teams throughout the business. However, that doesn't mean every team should be able to access every data point. In addition to basic role-based access restrictions, deploy your security data lake on a cloud data platform that offers these security capabilities:

- **Fine-grained access control** that allows database administrators to apply security constraints and rules to certain parts of each table, such as at the row level and column level.
- **Geofencing** so that administrators can set up and enforce access restrictions based on the location of each user.
- **Secure views** to prevent access to highly sensitive information that most users don't need to see. This security technique allows you to selectively display some or all of the fields in a table.
- **Data masking** that hides personally identifiable information (PII) or other sensitive information from users that require some access but shouldn't see all the details.

Tapping into a Cybersecurity Ecosystem

As explained throughout this book, a cloud data platform allows you to expand beyond basic SIEM use cases. You can ingest, parse, normalize, integrate, and analyze security data for a broad range of needs using best-of-breed connected partner applications, all easily accessible within an associated data marketplace. You own all the data and can apply it to use cases beyond SIEM, such as delivering actionable vulnerability remediation guidance through BI dashboards and rightsizing permission policies for SaaS and IaaS environments.

Review your SaaS vendors, including your identity provider and endpoint detection solution. Many vendors delete activity logs, transaction records, and other details from their systems after a few days, weeks, or months. With a cloud data platform, you don't have to work around these data retention windows. You can adopt a strategy of storing all the data generated by your vendors in one location for as long as you need. This creates a reliable source of truth for analysts who know where to go for the information they need.

A security data lake that holds all vendor data enables reliable generation of key security metrics such as visibility coverage, SLA performance, mean time to detect (MTTD) and mean time to respond (MTTR). A consumption-based pricing model allows you to have all the data on hand while paying for only the compute resources you use, making the outcomes described here not just possible but also cost effective.

In a typical deployment of a connected application with a security data lake, the partner vendor hosts the application infrastructure and code, then loads normalized and enriched security data into your cloud data platform instance. The bidirectional integration of these joint solutions means you have near-limitless storage and compute resources to power queries from within the SaaS applications. This model uniquely enables you to speed up your application vendor's performance by scaling resources as needed for crunching petabyte-scale data sets.

When planning your connected application deployment, work with the SaaS vendor and the data platform provider to calculate the combined cost for the first year. Ensure the estimate you receive is significantly lower than the legacy solution you are replacing or augmenting. Total costs should be cut by at least half to justify the migration effort. According to research conducted by Snowflake's Value Engineering team, consumption-based pricing for a cloud data platform, in conjunction with a connected SIEM or XDR application, is consistently 50% to 80% less than the prevalent ingestion-based pricing plans available from popular SIEM vendors (see Figure 3-3).

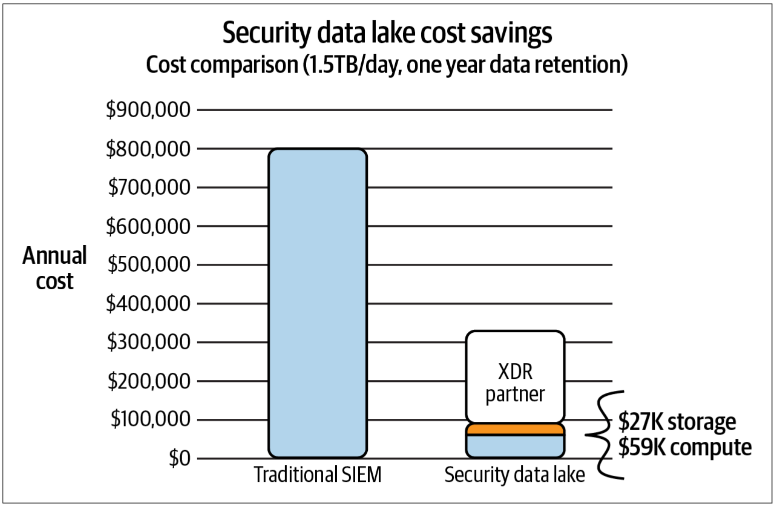


Figure 3-3. Consumption-based pricing allows customers to pay for actual resources used—per terabyte for storage and per second for compute (figures are based on 2020–2021 results)

Summary

Connected applications are SaaS solutions that separate code and data. The SaaS vendor maintains the application infrastructure and code while the customer manages the application's data within their own security data lake. Because the architecture rests on a scalable, cloud-built data platform, building and maintaining API integrations between the data lake and the connected apps is unnecessary. Each customer has its own environment within the cloud data platform where it can combine the data sets from multiple SaaS vendors with its own business data to create a unified, single source of truth for the entire organization.

Modern cloud data platforms power a broad vendor ecosystem that attracts best-of-breed security software vendors. Role-based access control, in conjunction with other security and governance mechanisms, ensures each user and each organization can access only the data they are explicitly permitted to see.

CHAPTER 4

Achieving Your Security Program Objectives

Bad actors are increasingly well funded, highly capable, and determined to leverage new technologies and paradigms to launch their attacks. In this constantly evolving threat landscape, building detections for every possible adversary or technique is nearly impossible. To maximize your defense posture, your security operation center must employ proven, repeatable processes for creating and maintaining threat detections in conjunction with continuous monitoring and testing to adapt them to real-world conditions. A security data lake enables you to achieve your security program objectives as part of a continuous process of improvement known as the Threat Detection Maturity Framework. This process yields more robust threat detections and greater alert fidelity by adhering to detection-as-code procedures.

Introducing the Threat Detection Maturity Framework

Incident response (IR) and threat hunting are separate but closely related activities. IR teams continually educate threat hunters about current attack patterns. Conversely, threat hunting teams share insights from their analytics so the incident response team learns more about what constitutes normal behavior in the environment.

To stay up to date on prevalent and emerging attack patterns, many security organizations adopt the MITRE ATT&CK matrix, an industry-standard framework for measuring threat detection maturity. This publicly available knowledge base tracks the tactics and techniques used by threat actors across the entire attack lifecycle. Created by a nonprofit organization for the United States government, the ATT&CK matrix helps security teams understand the motivations of adversaries and determine how their actions relate to specific classes of defenses.

Although the MITRE ATT&CK matrix is a good starting point, conscientious threat detection teams consider many additional factors outside of this matrix. A complete threat detection maturity framework should encompass five categories:

Processes
Development methodologies and workflows

Data, tools, and technology
Logs, data sources, integration logic, and documentation

Capabilities
Searches, analytic dashboards, and risk-scoring models

Coverage
Mapping of detections to threats, and prioritization of responses

People
A diverse and well-rounded SOC team

For each of these five categories, the framework defines the following three maturity levels:

Ad hoc
Initial rollout of security data, logic, and tools

Organized
Gradual adoption of best practices

Optimized
Well-defined procedures based on proven principles

For example, within the *processes* category, a team with an *ad hoc* level of maturity likely has no formalized development methodology, no defined inputs, only rudimentary detections, and no defined metrics.

As the team progresses to the *organized* stage, it establishes key development methodologies and workflows, defines important data inputs and detections, forges partnerships with connected application vendors, and collects essential metrics to gauge its progress.

Once the team reaches the *optimized* stage, the threat-detection effort includes a proven methodology and workflow, carefully delineated inputs for multiple detections, a well-defined threat lifecycle, and mature partnerships with connected application vendors. Metrics are continually collected and regularly presented to the CISO and other stakeholders accountable for the organization's risk posture.

The article "Threat Detection Maturity Framework" provides additional details on how to structure your threat detection efforts and progress along the threat detection maturity curve.

Embracing Detection-as-Code Principles

As security teams tasked with detecting and mitigating threats progress along the maturity curve, they should carefully consider how they develop, deploy, and maintain detection logic. Just as software engineers adhere to the DevOps lifecycle to build and maintain robust software applications, detection engineers should follow the *detection development lifecycle*, which is governed by *detection-as-code* principles.

DevOps thrives via peer review processes for developing, testing, and deploying new code. Detection-as-code applies the same concepts to the creation and maintenance of detection logic for identifying risks (proactively) and threats (reactively). The approach also extends "as code" to the collection of the security data and to the database schemas that define it. This gives security teams a structured way to make sense of security data at scale and to transform manual processes to automated ones. Without these rigorous, repeatable processes, your detections will generate false positives that may cause "alert fatigue." Your IR team won't be able to handle all the alerts, enabling some threats to progress from initial access events into full-blown breaches.

By taking lessons from the DevOps and DataOps disciplines, the *detection development lifecycle* allows security teams to leverage proven, repeatable processes for building, maintaining, and testing threat detections. It emphasizes reusable code, version control methods, peer review, and check-in/check-out procedures as analysts collaborate to create and maintain high-fidelity threat detections. Adhering to this lifecycle empowers SOC teams to develop robust security rules and monitor their performance in the environment.

The detection development lifecycle consists of the following six phases:

Requirements gathering

Collect relevant technical details from key stakeholders such as the primary goal of each detection, the systems these detections target, the risks and vulnerabilities they address, and the desired alerting methods (such as Slack, Jira, and other methods).

Design

Once work commences on a detection, the goal is converted into a detection strategy. Some security teams use standard detection frameworks such as the Palantir Alerting and Detection Strategy (ADS) framework, which also assists with creating documentation that defines the purpose and use of each detection.

Development

After a new detection's design has been completed, it is converted into code. Make sure every detection has a set of common fields and a link to your chosen detection framework so you can clearly define the goal of the detection in the code. Where possible, leverage out-of-the-box detections from connected applications.

Testing and deployment

Test each detection for accuracy, precision, and alert volume. Historical testing involves running the detection against past data. After testing is completed, detections are peer reviewed and managed in a version control system.

Monitoring

Continuously monitor the performance of deployed detections, review assumptions and gaps, and decommission detections that are no longer needed.

Continuous testing

This is how mature threat detection teams ensure each detection is accomplishing its intended goal. The output of continuous testing can be no action if the detection delivers appropriate alerts. It can also trigger a detection improvement request or a request for an entirely new detection.

Adhering to this lifecycle enhances the quality of your detections, encourages robust documentation of those detections, makes scaling your team easier, and serves as a solid foundation for developing program metrics, as described in the next section.

Improving Threat Detection Fidelity

The complexity of today's hybrid IT environments has led to an exponential growth in the number of alerts generated on a daily basis. Alerts arise not only from suspicious network behavior but also from routine internal events. For example, the finance department might roll out a new version of an enterprise software application that queries a large section of a key corporate database. As the software goes live, a sudden spike in CPU load on the database cluster might trigger multiple warnings, generating a flood of alerts spanning multiple systems. Other applications that rely on that cluster might experience memory deficiencies or latency issues.

These machine-generated alerts can be particularly challenging to isolate and identify because of the sheer volume of alert activity, or "noise," that can arise from a single incident. IR teams seek to correlate these alerts across multiple layers of the technology stack to determine if an incident represents a true threat or is simply the result of routine system maintenance. If the team uses a cloud data platform that can hold security data along with business data from these other IT activities, they can more easily bring in the necessary contextual information to eliminate or avoid false positives.

NOTE

Combining the holistic visibility of the security data lake with a detection-as-code approach reverses the traditional volume/noise trade-off. More data starts to mean less noise, not more.

For example, if a US-based employee suddenly appears to be logging in from another country known for malicious attacks, analysts may need to quickly determine whether the event constitutes an actual intrusion involving stolen credentials. In the traditional SIEM model, the security team likely would reach out to HR or the user's manager to find out if there's a good reason for the login location. This approach takes time and delays the incident response. A security data lake, as discussed in previous chapters, enables the security team to apply up-to-date contextual information (such as HR updates) to activity logs across multiple systems. With detections defined as code, the rule that resulted in the original alert would be modified to include the data sets the analysts used to investigate and dismiss the alert. Thus the next time this type of scenario occurs, the alert will be eliminated within the detection logic, avoiding the false positive. The fastest alerts to triage are the ones that were never triggered in the first place.

When HR data is stored in the security data lake and updates about employee status are monitored, that data can be correlated with the security data and instantly analyzed in time to prevent a potential breach or disregard a false positive. Some security teams also store issue tracking and project management data from Jira and other ticketing systems. Agile development teams use this data to track bugs, stories, epics, and other tasks. A high-fidelity detection can monitor when these tickets are created and approved. Being able to automatically log the ticket data removes manual processes from the security team.

By supporting structured, semistructured, and unstructured data types, a security data lake can not only detect many attacks but also support all the types of data that help you contextualize each incident to determine the root cause.

As detection rule sets evolve through the continuous improvement process, their accuracy reaches the point where IR teams can rely on the rule books for each alert. This maturity makes security orchestration, automation, and response (SOAR) activities more meaningful. If you've ever been disappointed by lackluster gains from SOAR, it was probably because low-fidelity detections required a human in the loop for most alerts. The combination of a security data lake and detection-as-code principles allows a SOAR program to achieve its full potential. Accurate and actionable detections trigger automated playbooks to stop a security breach or minimize its impact through actions such as issuing a security challenge or temporarily isolating compromised systems. It takes a lot of confidence to isolate a server in production, but a mature threat detection program can make this possible. Security data lakes also support advanced threat detections using data science techniques already employed in domains such as fraud detection, setting a near-limitless path towards improving detection maturity.

Preparing for Breach Response

In the event of a breach, security analysts must be able to study months' or even years' worth of data. These investigations may be performed directly within the security data lake using built-in SQL worksheets or via a purpose-built investigation interface within a connected application. Either way, results must arrive fast. This requires a single queryable repository for event data. The events that constitute a single incident may appear in one data set or be spread across many. These various events can be close in time or months apart. A key advantage to the security data lake architecture is that it eliminates the need for "rehydrating" or "replaying" events from cold storage. Data pipeline solutions that rely on data restoration often underplay the complexity and delay introduced within that approach.

Responding to a diverse set of threats requires a diverse set of data. A security data lake allows for long retention windows and can apply a consistent schema across sources from the entire IT infrastructure, both cloud-based and on-premises. This approach is much easier than working in a traditional siloed landscape, where the security team must investigate alerts and events console by console, API by API. Instead, the team can prepare for breach response by normalizing events and modeling unified views for assets, users, vulnerabilities, and other variables. This dramatically simplifies IR procedures when the team must respond quickly. The security data lake model offers the opportunity to prepare for fast IR in collaboration with the data analytics team. Connected applications handle most of the prep with prebuilt code and data models.

Measuring Alert Quality with KPIs

Once you identify your most critical log sources, such as data from firewalls, servers, and network devices, you can focus on improving the quality of the detections that operate against those sources. It's important to constantly scrutinize your threat detection workflows and incident response systems with an eye for continuous improvement. Which data sources yield the most alerts? Which alerts are the noisiest? Which detections yield the most false positives? Which data is the most critical? How strong are your rules?

With a security data lake, all your data is in one place, and all authorized users can access it via BI tools and data science models. BI dashboards can display current metrics such as the volume of phishing emails, the number of incidents, and the severity of incidents. If information about security and support tickets is stored in the security data lake, the dashboards can show the mean time to detect (MTTD) as tickets are escalated and mean time to respond (MTTR) as tickets are closed.

All these inquiries are made possible within a centralized platform where the team can query all data and insights simultaneously. Insights can be stored back into the data lake for future analysis. Mature security teams go beyond merely gathering lots of data into an analytics platform. Tremendous efficiency can be gained from metrics and key performance indicators (KPIs) that track the effectiveness of your cybersecurity efforts and enable data-driven decision making for future projects and initiatives.

Applying Data Science to Threat Hunting

Data science involves studying, processing, and extracting insights from a given set of information, such as the security data, log data, and contextual data sources you have identified as critical to your cybersecurity efforts. Data scientists create machine learning (ML) models that reveal trends and patterns in these data sets. For example, they might develop algorithms that identify the likelihood that certain types of devices, user profiles, or portions of a data set will be targeted during an attempted hack, or which types of network switches serve as the most common port of entry for denial-of-service attacks. Algorithms rank and score activity data to flag anomalies that may indicate suspicious behavior, maximizing the efficiency of security teams. Understanding these probabilities allows the security team to predict how potential attacks might unfold in the future.

Data scientists develop code using ML notebooks such as Jupyter and Zeppelin, as well as with high-level languages such as Python, Java, and Scala. If your security data lake supports these languages and notebooks, you can more easily enlist data scientists to conduct these investigations.

The detection development lifecycle, described in the preceding section, helps guide data science projects by enforcing regular procedures to develop ML models, review the code, run detections in test mode, study the true-positive and false-positive rates, and store the finished models in an alert library. The output of data science models then can be fed back into traditional business intelligence decision-making processes.

Business analysts from the cybersecurity and data teams can utilize SQL-based tools to view the results of investigations through self-service dashboards. This is an opportunity to gain more insight, and more value, from your security data.

Collaboration with data scientists is much easier when your organization standardizes on the same data platform. Skilled statisticians and ML engineers can get involved in cybersecurity investigations, just as they collaborate with other corporate domains such as marketing, sales, and finance.

Of course, data scientists need powerful compute resources to process and prepare the data before they can feed it into ML libraries and tools. The more data points they can collect, the more accurate their analyses will be. A cloud data platform allows them to easily access, collect, and organize data from a variety of sources and formats. The best cloud data platforms can scale compute and storage capacity separately and near infinitely, and they offer usage-based pricing, so you only pay for compute by the second. This cost model allows data scientists to ingest and process massive amounts of data at a reasonable cost.

Summary

To break down the data silos and enable analytics on a scale that can accommodate today's nonstop network activity, invest in a cloud data platform that can handle a broad set of use cases, including a security data lake, and work with a very high volume of data. Security teams can use this platform as a foundation to progress on the threat detection maturity framework and follow detection-as-code principles, including the following:

- Agile development of detections throughout the continuous loop of testing, debugging, deployment, and production
- Continuous integration/continuous delivery (CI/CD) of data pipelines and models for fast and reliable detection and response
- Automated testing and quality assurance (QA) for rules, especially important as upstream data sources change over time
- Versioning and change management for detection code
- Promotion, reuse, and automation of data models, detections, and other artifacts

By moving your data sets to a security data lake, you can reduce traditional SIEM license fees and operational overhead. You can use one system to analyze data from a huge variety of sources. You can store many types of data—including logs, user credentials, asset details, findings, and metrics—in one central place and use the same sets of data for multiple security initiatives. Collected data can be stored in the security data lake for however long you want, eliminating complex storage tiers and rehydration overhead. Anytime you want to search that data, you can do so easily via your connected security applications of choice.

A modern cybersecurity strategy begins with a security data lake and its rich ecosystem of security solutions and data providers equipped to handle the vastly expanding threat landscape.

About the Author

Based in Santa Barbara, California, **David Baum** is a freelance writer and marketing consultant who writes about innovative businesses, emerging technologies, and compelling lifestyles for a variety of print and online media.