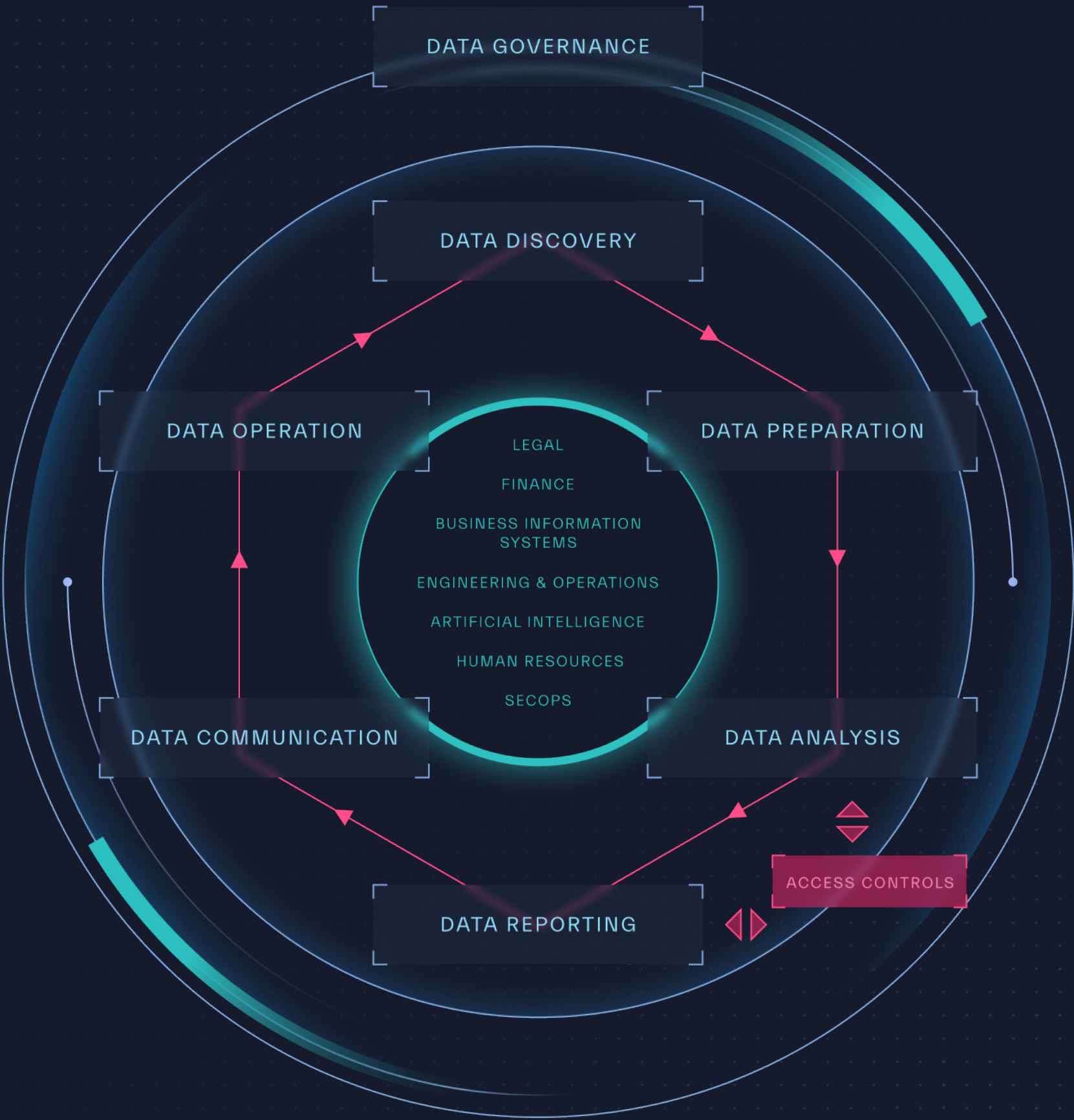


Applied Security Data Strategy: A Leader's Guide

A step-by-step guide to building a robust security data strategy, informed by industry experts



FEATURING CONTRIBUTIONS BY:

- Justin Borland Aqsa Taylor Alan Czarnecki Alex Gilelach Ryan Moon
- Matt Carothers Paul Keim Don Mallory Greg Olmstead Jon Oltsik

Contents

03	Acknowledgments
04	About the Authors
14	Preface
15	Data: The New Oil, Refining the Future
29	Data Discovery
43	Data Collection and Ingestion
49	Data Processing
58	Data Storage
65	Data Analysis
73	Data Reporting
79	Data Governance and Security
86	Understanding Common Data Platforms and Tools
98	Time to Build Your Own Security Data Fortress
108	Appendices
112	What Readers Are Saying

Acknowledgments

Applied Security Data Strategy: A Leader's Guide, is a collaboration between several security professionals from a wide variety of backgrounds, including Financial Services, Health Care, Retail, Telecommunications, Manufacturing, and Government.

All collaborators have a proven track record of creating, implementing, and maintaining various facets of their organization's data program.

Special thanks to:

Justin Borland (Abstract Security)

Aqsa Taylor (Abstract Security)

Alan Czarnecki

Alex Gilelach

Ryan Moon

Matt Carothers

Paul Keim

Don Mallory

Greg Olmstead

Jon Oltsik

About the Authors

Justin Borland



Justin Borland is the Director of Threat Engineering at Abstract Security. A proven technical leader in the security industry, Justin started his career with a Canadian Secret clearance while still in college. After graduating, he spent the next decade building custom packet capture systems, intrusion detection systems, logging systems, and DFIR tooling for large organizations.



Justin established and ran the Countermeasures team at Equifax. The team was responsible for building and maintaining the fleet of Moloch PCAP/IDS, IPS, and hundreds of other systems using petabytes of data. His team was also responsible for discovering the 2017 data breach.

Justin was called into the Senate, while on paternity leave, in 2018. At Barclays, Justin was part of the Global Hunt team and helped develop and mature many Threat Hunting capabilities and processes, especially related to malware analysis and DFIR. Justin led both the Threat Detection & Response and IT Ops functions at Unqork, a no-code/codeless-as-a-service company.

About the Authors

Aqsa Taylor



Aqsa Taylor is the Senior Director of Product Marketing at Abstract Security. She is a published author of the book “Process Mining: The Security Angle” and has extensive Product Management experience, having built some of the industry's most well-known cybersecurity platforms, such as Twistlock (acquired by Palo Alto Networks), Prisma Cloud Workload Security by Palo Alto Networks, and Gutsy (the industry's first AI-driven process mining platform).

Throughout her career, Aqsa has actively engaged with and consulted CISOs and information security leaders from both public and private sectors, including more than 45% of Fortune 100 organizations, on best cloud security practices.

She is also known for her cybersecurity evangelism, speaking at conferences, podcasts, and webinars about industry trends and security best practices. Her technical articles have been recognized by global magazines and media outlets such as Dark Reading, Security Boulevard, Unite.ai and others.

About the Authors

Alan Czarnecki



Alan Czarnecki is an experienced security operations professional with extensive expertise in threat hunting, incident response, and anti-phishing strategies. He spent time with a large Midwest insurance company, where he developed unique methods for anomaly detection and incident response.



Alan holds a Master of Science in Electrical Engineering from The Ohio State University and certifications including CISSP and GIAC Certified Incident Handler. His contributions to the field include professional presentations at events like FS-ISAC and A Conference on Defense. Alan's career is defined by a commitment to developing advanced solutions that optimize network visibility and safeguard organizational assets.

About the Authors

Alex Gilelach



With over a decade of leadership in Cybersecurity and Technology, Alex Gilelach is an expert in building and running cybersecurity and technology programs for large enterprises and financial institutions. Most recently, as the leader for Cyber Threat Intelligence at a large US financial institution, Alex developed the CTI program, integrating it into the Cyber Operations function to enable proactive threat defense.

Prior to CTI, Alex served as the Director of Security Architecture, responsible for deploying and operationalizing Cybersecurity tools for a multinational financial institution.

Alex also brings expertise in enterprise and application architecture, solution, and software development on distributed enterprise systems. In his prior roles Alex has led systems performance management and automation initiatives, enhanced operational efficiency and reduced risks through scalable, automated solutions.

Alex holds a Bachelor of Science in Computer Information Systems. He is engaged with multiple industry working groups for threat intelligence sharing and is fluent in eastern European languages.

About the Authors

Ryan Moon



Ryan C. Moon, known as _moon most places, is a seasoned professional with 20 years of experience in the financial services and fintech industries, having worked in security innovation and countermeasures engineering for two of the world's largest financial services providers. He played a crucial role in developing one of the most successful fraud actor targeting systems for the financial services industry.



Ryan co-founded one of the largest private cyber intelligence communities globally and has been providing free and mixed-source community-facing intelligence tools for a decade. Currently, Ryan works in a skunkworks start-up team within an established fintech, focusing on cybersecurity and incident response for large investment wealth platforms, financial services providers, and the automotive industry.

About the Authors

Matt Carothers



Matt Carothers is a Sagittarius. He enjoys sunsets, long hikes in the mountains, and intrusion detection. After studying Computer Science at the University of Oklahoma, he accepted a position with Cox Communications in 2001 under the leadership of renowned thought leader and virtuoso bass player William “Wild Bill” Beesley, who asked to be credited in this bio. There Matt formed Cox's first customer safety department, which he led for several years, and today he serves as Cox’s Senior Principal Security Architect. On weekends you can usually find him crashing radio-controlled planes and 3D printing replacement parts for them.



About the Authors

Paul Keim



Paul Keim is a Senior Security Architect at Cox Communications. After getting his Masters in Computer Engineering from the University of Florida, Paul entered the IT Security consulting world. After determining management was not for him, he left to join Cox Communications, leading their Application Security team.



In the past few years, he’s shifted to work in Defense Engineering and Security Operations teams, providing support for logging infrastructure and its downstream tooling. After much nudging from Matt Carothers, Paul has begun working on white papers and sharing his experience with the greater security community.

About the Authors

Don Mallory



Don Mallory is a graduate of the Mohawk College Computer Electronics Engineering Technology program and holds of a number of certifications with over 30 years of experience in enterprise IT, primarily in critical infrastructure, specializing in operations, large scale data storage, disaster recovery, and security for critical infrastructure across multiple continents.



Currently, Don is a Senior Security Analyst in the healthcare sector. He has spoken as a keynote and invited speaker at multiple conferences nationally in Canada on cybersecurity and privacy topics. Don has also been involved in various volunteer activities including the Canadian Collegiate Cyber Exercise (C3X) as a builder and mentor, co-organizer of Hak4Kidz Toronto, and the Latow Photographer's Guild at the Art Gallery of Burlington, where he teaches traditional wet darkroom photography.

About the Authors

Greg Olmstead



Greg Olmstead is a recognized expert in cloud computing, cybersecurity, and data architecture with a proven track record of delivering innovative solutions across the financial, energy, and fintech sectors. As a Cloud Security Architect at Bloomberg, Greg has led large-scale initiatives to enhance cloud security, streamline infrastructure automation, and improve compliance across global cloud environments.

With extensive hands-on experience in platforms like AWS, Azure, Google Cloud, and Kubernetes, Greg specializes in creating secure-by-design solutions, including hybrid encryption systems, developer-centric security frameworks, and robust network security patterns. His work has enabled organizations to securely scale their cloud infrastructure while minimizing risk and optimizing performance.

A leader in cloud security innovation, Greg holds one issued patent and has another pending. He has contributed to open-source projects, developed proprietary threat detection systems, and presented his work at multiple technical conferences. Whether designing advanced infrastructure-as-code frameworks or embedding security into every layer of cloud operations, Greg's mission is to help organizations harness the power of data while safeguarding it against evolving threats.

About the Authors

Jon Oltsik



Jon Oltsik, is a former distinguished analyst and fellow at the Enterprise Strategy Group and the founder of its cybersecurity practice in 2003. Jon remains active in the cybersecurity community; writing, consulting, and working with CISOs, cybersecurity practitioners, threat intelligence analysts, and security technology vendors.



Preface

Applied Security Data Strategy: A Leader's Guide is designed to help organizations of all sizes and sectors—including finance, education, health, entertainment, government and SaaS services. The goal of this book is to help people and organizations evaluate, assess and mature their current data strategy maturity and find step by step instructions to improve it. It is useful for every organization regardless of their current maturity state.

More than just a theoretical guide, the **Applied Security Data Strategy: A Leader's Guide** is a practical toolkit for building a modern security data program. We delve into the building blocks of a successful and future proof data strategy that is based off the real-life experiences of experts in the field. This resource leaves no stone unturned in addressing the daily problems data security professionals face in their operations.

Inside, you'll discover:

- 1

The key pillars of a successful security data strategy:

From data discovery and collection to reporting and governance, this eBook covers every stage of the security data lifecycle.
- 2

A maturity model for each pillar:

After every chapter, assess your organization's current capabilities and identify areas for improvement.
- 3

Real-world examples and use cases:

Learn from our experts directly and see how the concepts outlined in this resource are applied in different industries.
- 4

Step-by-step instructions and practical advice:

Gain actionable insights and guidance on how to implement, adapt, and optimize each component of your security data strategy.
- 5

A comprehensive understanding of security data strategy:

Learn how this model can address the daily challenges you face, and how to build a strategy that works for you.

Chapter 1

Data: The New Oil, Refining the Future



Data: The New Oil, Refining the Future

Your identity, your assets, your bank accounts, your entire existence is stored as a digital record of data somewhere. Any compromise of an organization dealing with data today has huge ripple effects, not just on the organization, but every affected party, including the country the event took place in. It is no surprise that governments have taken it upon themselves to enforce strict regulations around data privacy, compliance, governance, and reporting, on top of industry and organization specific policies.

No matter what industry you work in, you deal with data management in some way. Having an effective data management strategy with a security lens is not just a ‘nice to have’ but a strict requirement. With the growing digital transformation leading to growing data volumes (Figure 1.1) and regulations getting stricter by the day, not getting this right isn’t just risky, it’s a huge liability. Data management is something you simply can’t afford to overlook.

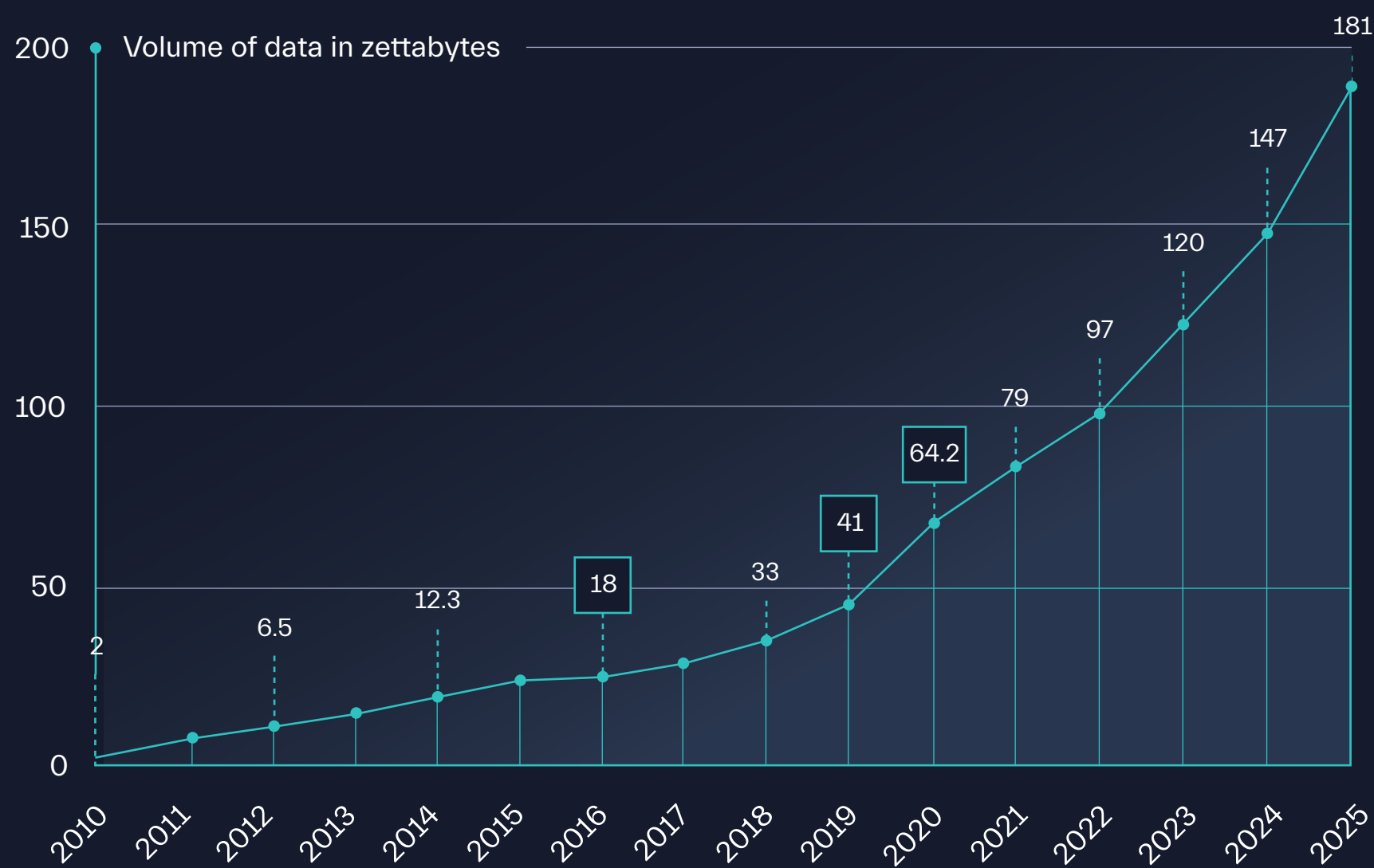


Figure 1.1: Increasing volumes of data worldwide
Source: [Statistica.com](https://www.statista.com)

Security Operations: It's All About the Data

In just the past two years, managing security data has become dramatically more difficult, according to nearly half of cybersecurity and IT professionals surveyed by the [Informa TechTarget's Enterprise Strategy Group](#). The modern security landscape is drowning in data with every endpoint, server, application, and network device generating a torrent of logs and alerts. Most organizations are stuck in a rat race, trying to collect it all. Even though having a constant influx of information is indispensable in detecting and responding to potential threats, it presents a significant challenge for security teams.

The modern security landscape is drowning in data with every endpoint, server, application, and network device generating a torrent of logs and alerts.

This problem stems from an organizations' attack surfaces expanding due to cloud adoption, remote work, and increasing number of Internet of Things (IoT) devices, compounded by an ever evolving threat landscape. As more devices connect to networks, it inevitably attracts attempts from malicious attackers who will use a wide range of tactics to gain illegal access, making it even harder for security teams to sift through noise and identify real threats.

Furthermore, [Enterprise Strategy Group's data indicates](#) that two-thirds of organizations have at least four different repositories for security data, which include Security Information and Event Management (SIEM), Extended Detection and Response (XDR), Network Detection and Response (NDR), Endpoint Detection and Response (EDR) platforms, vulnerability scanners, and more. This creates fragmented data sources, visibility gaps, performance issues with analysis and queries, among other complications. As such, even though security teams may be armed with the tools they need to safeguard critical data, they may be lacking an effective data strategy to successfully manage and analyze the vast amounts of data at their disposal.

The Impact of Underserved Security Data Management

Fragmented data sources and visibility gaps.

Some data sources are in an obscure format, some are incomplete, and some remain outside the purview of the security operations team. Investigating security incidents depends upon transparent and frictionless access to all data sources. Poor security data management can greatly inhibit the Security Operations Center (SOC) team's ability to detect and respond to cyberattacks in a timely manner.

Performance issues associated with analysis and queries.

SOC teams often collect and process terabytes of data, but much of this data is used for regulatory compliance making it unnecessary for threat detection and risk mitigation. This results in massive data repositories for security operations, creating scalability and performance issues. When a level 2 SOC analyst performs a complex SIEM query, it can take minutes or hours to execute. Obviously, delays like this can be the difference between rapid containment and further damage.

Vendor lock-in and SIEM migration challenges.

To optimize security operations, Chief Information Security Officers (CISOs) want the flexibility to add new analytics tools or migrate from legacy to cutting-edge SIEM solutions. Ineffective security data management creates hurdles toward these goals by constraining SOC teams to specific data formats, query languages, and detection rule sets. This makes migration extremely difficult. The burden of migration costs and effort is so great that it often prevents SOC managers from implementing innovative security technologies, even when those technologies offer clear benefits.

Difficulty in maintaining compliance.

Auditors and risk officers need timely access to security data, preferably delivered in well-organized regulatory compliance reports. With data scattered in multiple formats and repositories, compliance reporting can take far more time and effort than it should. When SOC teams struggle to piece together data for compliance reporting, they expend resources which could otherwise be used for security monitoring and threat detection.

Data storage dilemmas.

Managing terabytes of distributed security data results in a compromise between speed of retrieval and storage costs, with most organizations struggling with endpoint security data (Figure 1.2). According to Enterprise Strategy Group’s research, 96% of security professionals agree that their organizations would benefit if they collected, processed, and analyzed more security data, but cost and resource limitations preclude them from doing so.

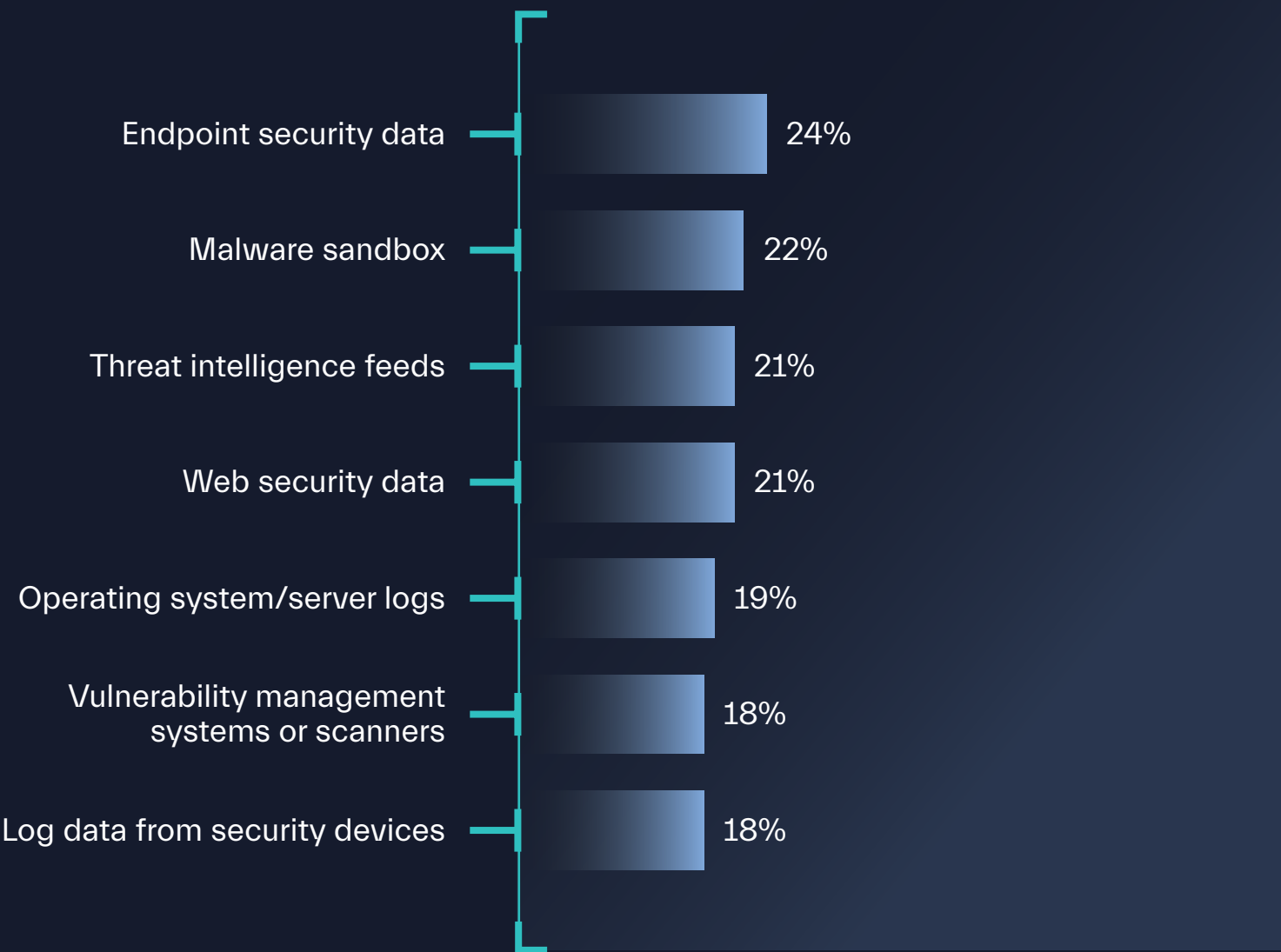


Figure 1.2: SecOps storage data survey. Source: [Informa TechTarget’s Enterprise Security Group Research Report, The Triad of Security Operations Infrastructure: XDR, SIEM, and MDR, June 2024.](#)

Security Operations Continue to Grow More Difficult

While data management is a key component of strong security, security operations themselves have become increasingly difficult. According to recent research from Enterprise Strategy Group, nearly half (45%) of security professionals surveyed claim that security operations are much more difficult for their organizations today than they were two years ago for several reasons including (Figure 1.3):

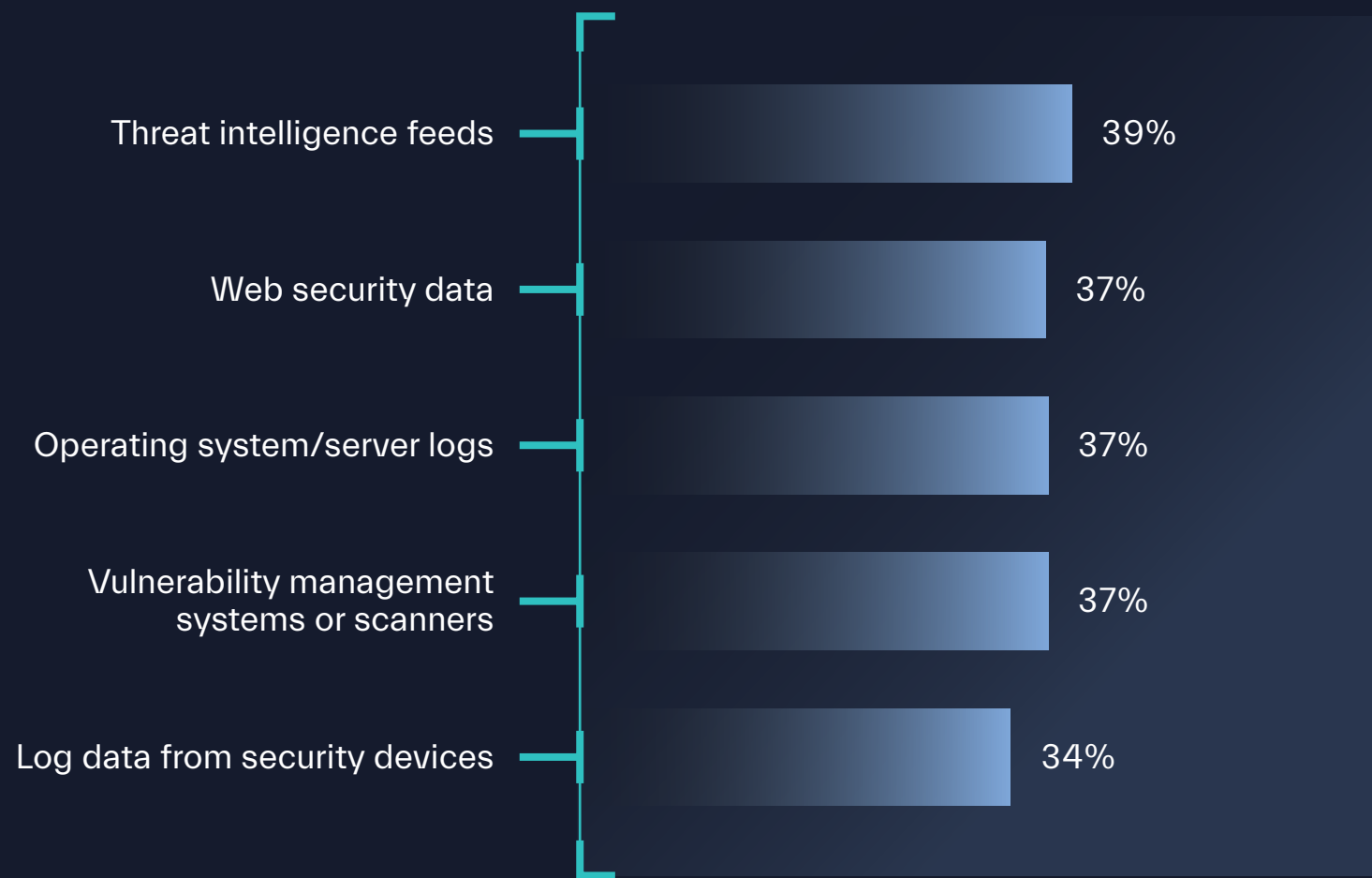


Figure 1.3 Survey of threat landscape incidents. Source: [Informa TechTarget’s Enterprise Security Group Research Report, The Triad of Security Operations Infrastructure: XDR, SIEM, and MDR, June 2024.](#)

A growing attack surface.

New cloud-based workloads, applications, Application Programming Interfaces (API)s, and third-party IT connections make the external attack surface extremely difficult to monitor and manage. Doing so requires continuous cooperation between security, IT, and software development teams to analyze a multitude of data sources for detecting, prioritizing, and remediating vulnerabilities as quickly as possible.

A rapidly evolving and changing threat landscape.

According to threat intelligence from Check Point research, global cyber-attacks increased by 30% in Q2 2024, reaching 1,636 weekly attacks per organization. Cyber-attacks are especially focused on industries such as financial services, technology, healthcare, and public sector agencies. To keep up, security professionals must collect, process, analyze, and act on numerous commercial and open-source threat intelligence feeds.

Increasing volume and complexity of security alerts.

In 2020, Forrester Research reported that security teams dealt with an average of 11,000 alerts per day, and that 28% of these alerts (3,080) were never addressed due to staffing and skills shortages. Given the increase in the attack surface, cyber-threats, and data sources, it is safe to assume that these numbers have increased significantly over the past five years.

A deluge of security data.

Consistent with the description above, nearly two in five (37%) of organizations collect and process more security log data today than they did two years ago, making security operations more difficult. This is an ongoing trend at most enterprises due to the previously described factors (i.e., a growing attack surface, an evolving and changing threat landscape, etc.). CISOs should expect even more security data with further adoption of cloud native applications, AI implementation, and digital transformation.

What is Security Data Strategy?

A data strategy is the foundation for all data initiatives. It is a comprehensive guiding plan that defines how data should be collected, stored, and used to support broader business goals. A security data strategy is a data strategy developed with security in mind throughout the entire data lifecycle from assessment, through use, and completing with deprovision and destruction.

The purpose of this book is to enable organizations to assess and implement a data strategy with a security focus, ensuring that every pillar of the strategy accounts for data security and governance.



Breaking Down Common Myths

Moving forward with a data-centric security approach means understanding and overcoming some common myths:

MYTH 1: THE LOG LEMMINGS DILEMMA

“Log everything.” Sounds like good advice, but without caution it can go wrong in many ways for security. Logging everything is a huge task, but that is not the problem. The problem, surprisingly, becomes visibility.

You might think that is an odd conclusion, but let’s break it down. An example log entry would have the following levels:



INFO Level

Informational level - indicating an event happened. Most of this would not be useful for security.

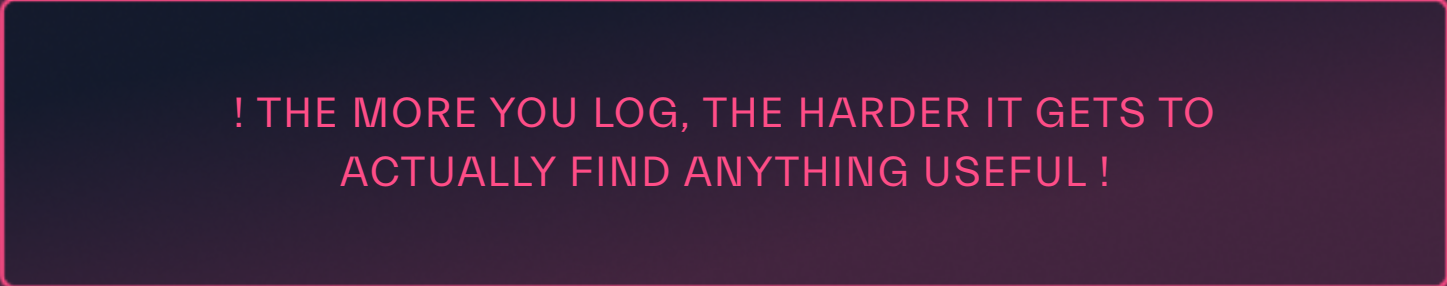
DEBUG Level

System process level events are useful for debugging should something go wrong. Somewhat useful if there is an impact on the system because of a security incident.

ERROR Level

When something goes wrong. Not all errors are security related; it could be a system failure completely unrelated to any real threat.

When you collect and store everything without preset filters or tuning, you end up with a lot of noise and low-quality data mixed in with your high-priority and critical alerts. So, when something goes wrong and you need to search for those high-priority alerts, you're searching through all that noise, which slows down the process and makes the results harder to interpret.



**! THE MORE YOU LOG, THE HARDER IT GETS TO
ACTUALLY FIND ANYTHING USEFUL !**

Over time, logging so much data just creates a noisy mess where important security details get lost. There is a direct cost to logging itself, but the exorbitant cost of the amalgamation of other factors such as the direct material cost of licenses, processing and storage of logs drives the total cost of ownership to unmanageable levels. Ultimately, it is the effort required by the Security Operations (SecOps) team to sift through all that data in hopes of finding critical security events.

MYTH 2: COLLECTING ONLY SECURITY EVENTS IS ENOUGH

Yes, this may sound like a contradiction to myth #1, but finding the right balance between the two is key. Logging only security alerts without context can cause serious issues when conducting a Root Cause Analysis (RCA) or trying to report the chain of events leading up to an attack. Identifying which information would be useful in a security investigation is key to understanding what entire logs versus specific fields within those logs is required. In addition to security alerts, you may need to collect other supporting evidence from event data, such as audit logs, to further correlate and analyze the threats.

MYTH 3: SECURITY IS A STEP IN THE PROCESS

“Data Governance and Security” is often defined as a step in the overall data strategy. Our current threat landscape has proven that if anything, security and governance should be an over-arching umbrella over the entire data strategy. All aspects of data interaction must be governed in accordance with applicable laws, regulations, industry standards, and internal policies. Failing to adhere to these will result in legal, financial, and reputational risks.

Building Toward a Better Security Data Strategy

This book introduces the Security Data Strategy Maturity Model, a collaborative framework developed by security professionals across diverse industries. It serves as a roadmap for organizations to assess their current data capabilities and identify areas for improvement. In doing so, it will outline key components of a successful data strategy—including data collection, ingestion, processing, storage, analysis, governance, monitoring, and reporting. By understanding the maturity levels within each component, security teams can chart a course toward data-driven security excellence.

Using a mature data strategy, security teams can create a lasting framework to:

1. Gain Clarity

Instead of being overwhelmed by a constant barrage of data and alerts, teams can prioritize high-fidelity signals and respond to real threats faster. This translates to improved analyst capability across various security functions, from SecOps and threat intelligence to Digital Forensics and Incident Response (DFIR) and Site Reliability Engineering (SRE).

1. Improve Efficiency

Optimize resources and accelerate response times. Increase tech deployment and adoption velocity with custom or incident-based tools and improve use of incident-response-as-code and forensics-as-code.

1. Enhance Threat Detection

Identify and respond to sophisticated attacks more effectively. By correlating data from various sources and applying advanced analytics, teams can detect complex attack patterns that might otherwise go unnoticed.

1. Enable Proactive Security

Anticipate future threats instead of only reacting to them. Leverage historical data and threat intelligence to identify vulnerabilities and proactively strengthen security posture. This includes having the right data on demand and knowing what to collect now versus after a security incident.

1. Move Beyond Achieving Compliance

Exceed regulatory requirements and demonstrate accountability to stakeholders. A strong data strategy helps companies comply with data privacy regulations and industry standards, reducing the risk of legal and financial penalties.

Pillars of a Security Data Strategy

In the following chapters we will dive deeper into each element of an effective security data strategy (Figure 1.4) and the way to assess an organization’s maturity level in each stage.

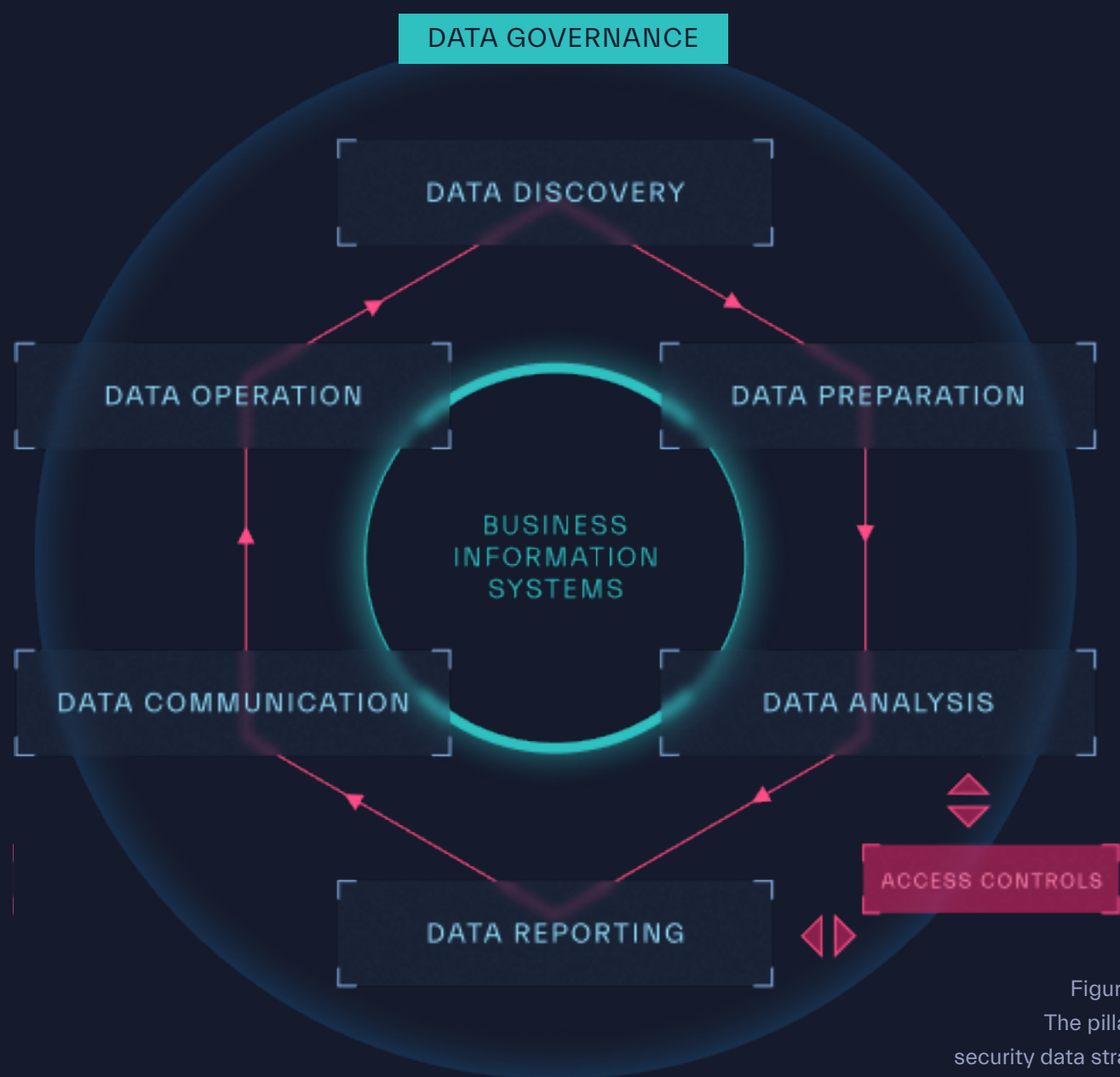


Figure 1.4:
The pillars of
security data strategy.
Source: [Abstract Security](#)

The core elements of a security focused data strategy are:

1. Data Discovery
2. Data Collection and Ingestion
3. Data Processing
4. Data Storage
5. Data Analysis
6. Data Reporting
7. Data Governance

The foundation for all pillars is Data Governance. Each of the following chapters will explore individual pillars in greater detail and provide a maturity assessment model to help you analyze your current data strategy maturity and identify what is needed for improvement.

Chapter 2

Data Discovery



Data Discovery

It is highly likely that teams are swimming in a sea of information. Whether it’s customer-generated data, financial transactions, or sensor readings, data is constantly being generated at an unprecedented rate. However, as useful as data is, it is only valuable if it can be effectively harnessed. That is where data discovery comes in (Figure 2.1). In this chapter we will dive deep into the key phases in data discovery and its maturity assessment model.



Fig 2.1 Data Discovery benefits Source: [Polymerhq.io](https://polymerhq.io)

What is Data Discovery?

Data discovery is the process of identifying, cataloging, and analyzing data assets within an organization. It provides a comprehensive view of the entire data landscape, enabling organizations to make informed decisions, improve operations, and ensure compliance with regulations. Think of it as creating a detailed map of the organization’s data universe. This will help your teams ‘discover’ hidden treasures that can lead to valuable insights.

Data discovery initiatives are triggered by a variety of needs, which can be broadly classified in two ways: Tactical and Strategic (Figure 2.2).

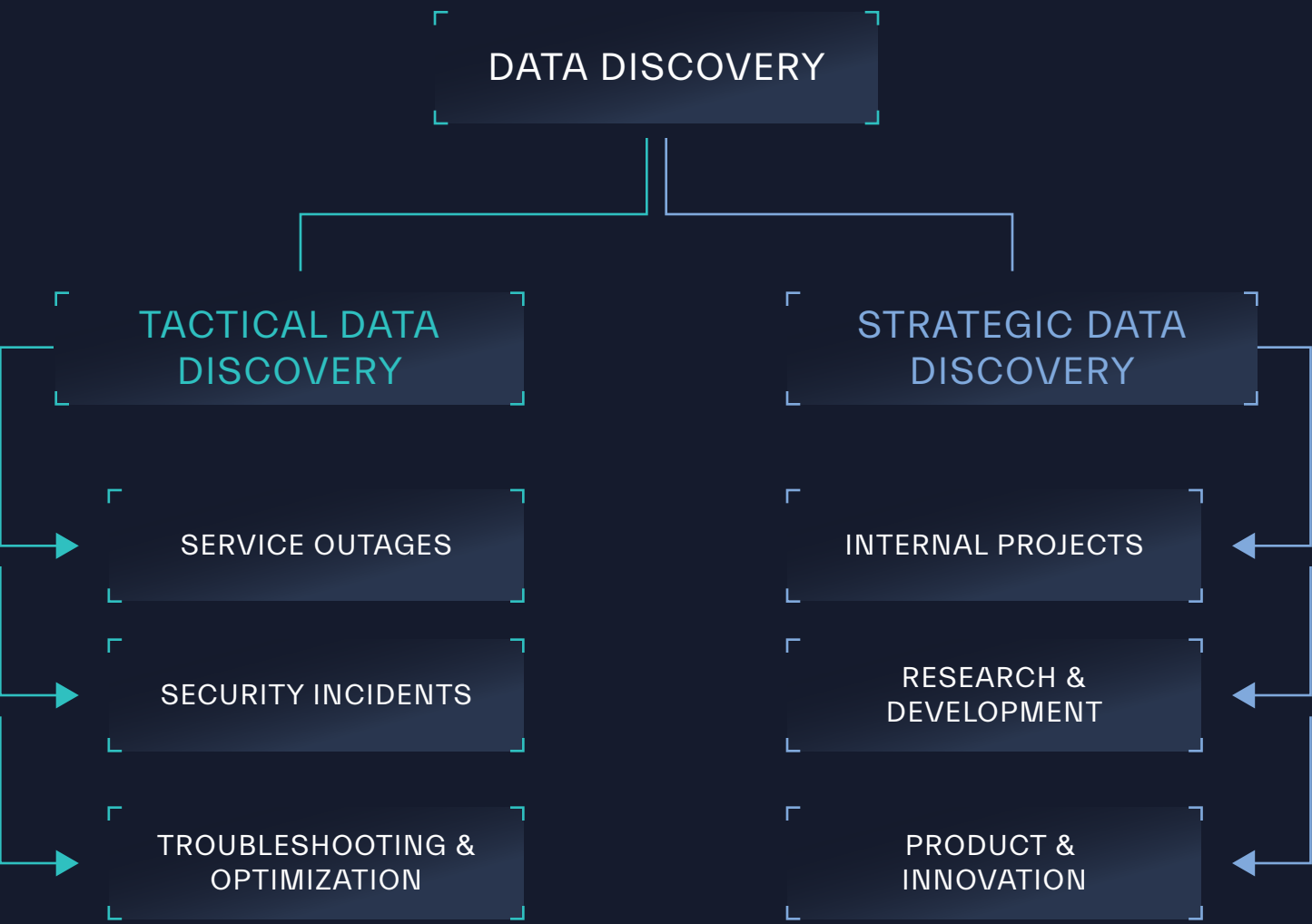


Figure 2.2 The needs of data discovery. Source: [Abstract Security](#)

Tactical Data Discovery

Tactical data discovery arises from urgent situations that demand immediate action. This often involves responding to events such as security incidents, where teams need to quickly identify the source of a breach, the affected systems, and the extent of data compromised. Similarly, during DFIR investigations, data discovery plays a crucial role in uncovering evidence and understanding the attacker's methods.

Compliance and regulatory requirements also frequently necessitate tactical data discovery. For example, if an organization receives a data subject access request under General Data Protection Regulation (GDPR) it needs to quickly locate and provide all personal data related to that individual. Other tactical needs might include troubleshooting performance issues, investigating service outages, supporting mergers and acquisitions, or responding to financial and regulatory audits.

**! TACTICAL DATA DISCOVERY IS DERIVED FROM
AN URGENT RESPONSE TO A TRIGGER !**

Strategic Requirements for Data Discovery

Strategic data discovery is a more planned and organized process which is often tied to ongoing projects and deployments. These initiatives are typically part of broader organizational goals, such as enabling new business units, preparing for a certification or audit, or improving operational efficiency.



**! STRATEGIC DATA DISCOVERY IS A PLANNED APPROACH,
AS OPPOSED TO A REACTIONARY COURSE !**

Similarly, as organizations increasingly rely on cloud services, data discovery helps map out data landscapes across various cloud environments, ensuring data security, compliance, and cost optimization. An example of strategic data discovery is when a company adopts a new customer relationship management (CRM) system. In this situation, data discovery becomes crucial to understand existing customer data and how it can be migrated to the new platform.

Strategic data discovery also plays a vital role in maximizing the value of new tools and technologies. When a business unit acquires a new analytics platform, data discovery helps identify relevant data sources and integrate them with the platform to generate meaningful insights.

Unexpected Discoveries

Sometimes data discovery can occur by chance. This often occurs during incident response or when exploring new parts of your IT infrastructure. For example, while investigating a server outage, IT professionals might stumble upon an obscure database containing valuable customer information. Such accidental discoveries can then inspire strategic data discovery plans, leading to unexpected benefits and opportunities.

organizations must continuously reassess their data discovery practices to ensure they are capturing all relevant data

It is also important to recognize that data landscapes are dynamic. New data sources will always emerge, old ones will eventually become obsolete, and data requirements will constantly evolve. As such, organizations must continuously reassess their data discovery practices to ensure they are capturing all relevant data while minimizing the collection of unnecessary information. This might involve revisiting previously excluded systems, updating data retention policies, or adapting to new data privacy regulations.

The Five Major Phases of the Data Discovery Process

Data discovery is a multifaceted process with distinct stages (Figure 2.3), each contributing to a comprehensive understanding of an organization’s data assets:



Figure 2.3
The stages of data discovery.
Source: [Abstract Security](#)

PHASE 1: SOURCE IDENTIFICATION

The first step in data discovery is identifying all potential sources of data within your organization. This can encompass a wide range of origins, from traditional databases and file systems to cloud storage, applications, social media platforms, and even IoT devices (Figure 2.4).

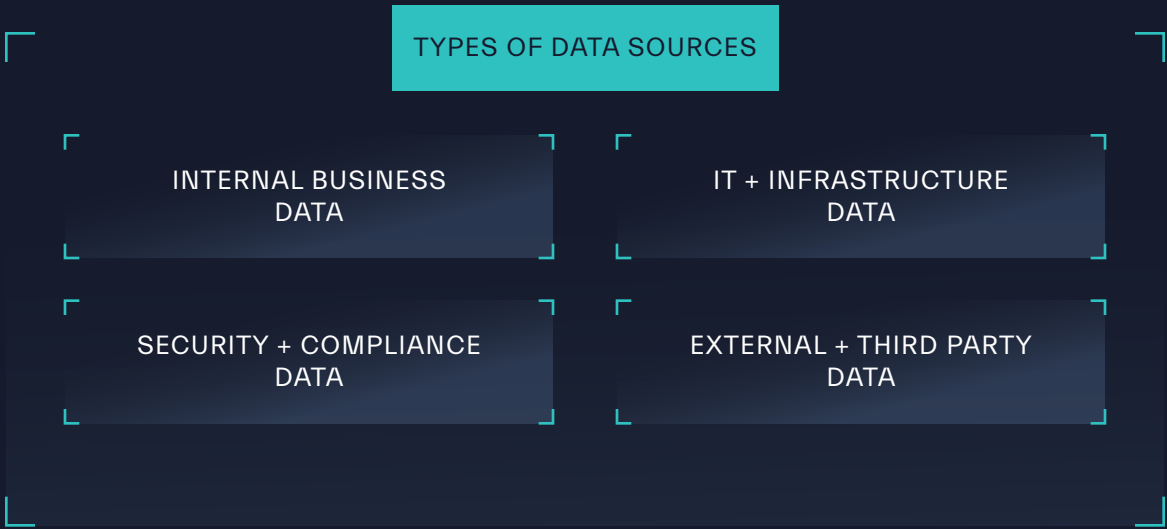


Figure 2.4 Types of data sources. Source: Abstract Security

For example, in a health care setting, data sources might include electronic health records, medical imaging systems, patient portals, wearable devices, and research databases. Meanwhile, in a manufacturing environment, data could reside in production line sensors, quality control systems, supply chain management software, and customer relationship management platforms.

Organizations must be comprehensive and thorough in the quest to understand the full scope of the data landscape.

! ONCE A CERTAIN TYPE OF SOURCE HAS BEEN IDENTIFIED, AN EXHAUSTIVE SEARCH SHOULD BE DONE TO ENSURE THAT ALL KNOWN SOURCE TYPES WITHIN THE DIGITAL ESTATE ARE TREATED THE SAME AS PER THEIR DATA REQUIREMENTS !

Just because a database or service has always been used for one kind of activity, doesn't mean that it is being used with all the required safeguards.

PHASE 2: INITIAL CATALOGING

Once your data sources are identified, the next step is to systematically catalog them (Figure 2.5). This involves creating a centralized inventory that documents key information about each data asset including details about the type of data (structured, unstructured, or semi-structured), location, format, sensitivity, and its owner (the department or individual responsible for the data).



Figure 2.5 Cataloging details for data sources. Source: Abstract Security

! EFFECTIVE CATALOGING INVOLVES CAPTURING METADATA, WHICH PROVIDES CONTEXT AND MEANING TO YOUR DATA !

Metadata captured during the cataloging activity can include information about the data’s origin, its creation data, its quality, and whether there are requirements for ingestion or safe access. However, it is equally important to perform data labeling so that all of the metadata you captured is easy to find, understand, and is used effectively (Figure 2.6).

Initiative Support	Policy
Executive Sponsorship	Scoped to Include: Data in all Forms Data in all Locations
Engage Stakeholders	Descriptions
	Clear Examples
	Impact of Loss / Disclosure
	Storage and Controls
	Transport and Handling
	Disposal, Declassification

Figure 2.6: Implementing Data Labeling. Source: E-Health Ontario

You also need to make sure that you’re properly labeling your data correctly for retention purposes, which includes dispositioning, disposing, and destroying data as needed. Every business needs to follow laws and regulations, and depending on your industry, there may be even more sensitivities you need to keep track of (Figure 2.7). For example, even if an employee leaves your organization, you still need to keep their HR record data on-hand for about three years. However, leaving it in your systems too long can potentially introduce different legal issues.

You should implement these kinds of controls from the start, as not only does this help your discovery process, but it can also inform other stages such as storage. Some pieces, like nuclear safety data, have to be available for approximately 150 years! With that in mind, does your sector have any unique retention requirements? If so, label them accordingly and then determine an efficient (and cost-effective) way to store them.

Type of Records	Retention Duration
HR	3 Years after Employee Departure
Financial	7 Years
Corporate	10 Years
Patents	20 Years
Patient Records	18–25 Years (Under 18), 25 Years (Over 18)
Engineering Drawings	Life of Subject + 15 Years
Nuclear Safety Data	30 Years Beyond Decommissioning (5+5+5+25+5+10+10+30=150 Years)

Figure 2.7: Cataloging for Retention. Source: oha.com

PHASE 3: DATA PROFILING AND SAMPLING

Data profiling delves deeper into the characteristics of your data. It involves analyzing the structure, content, and quality of each data source to understand its potential value and identify any potential issues. This might involve examining the distribution of values in a dataset, identifying data types and formats, and assessing data quality by looking for inconsistencies, missing values, or outliers.

Data sampling is often used during profiling, especially when dealing with large datasets. By analyzing a representative subset of the data, insights can be gained into the overall characteristics of the dataset without having to process the entire volume.

PHASE 4: PATTERN AND KEYWORD DISCOVERY

This phase focuses on identifying sensitive data that may require special handling. A systematic set of searches should be compiled and completed to ensure any data types that potentially may have legal, regulatory, or other governance requirements are identified (Figure 2.8).

This includes personal data, financial data, health information, and intellectual property such as trade secrets, patents, and copyrights. Additionally, any data that falls under regulations such as the GDPR, the Health Insurance Portability and Accountability Act (HIPAA), the California Consumer Privacy Act (CCPA), or other legislative frameworks may require auditing before continuing the discovery process to ensure compliance.

Defined for Regulated Data
Confidential + PII
Confidential + PHI
Confidential + PCI

Increased Granularity for Data Subject Rights
Confidential-Financial
Confidential-HR
Confidential-Legal
Confidential-Technical
Confidential-PII-Identifier
Confidential-PII-Identifying
Confidential-PII-Confirming

Figure 2.8: Sensitivity – Alternative Models.
Source: Don Mallory (Appendix M)

Organizations should use various techniques to identify sensitive data.

- Regular expressions (regex) search for specific data formats.
- Data dictionaries universally define meaning, format and data elements for the entire organization.
- Machine learning algorithms can be trained to identify sensitive data based on patterns and context.
- Samples of specific data can be used to identify exact data match elements.

PHASE 5: SYSTEM OWNERSHIP AND ENVIRONMENT MAPPING

The final phase of data discovery focuses on establishing clear ownership and accountability for each data source. This is crucial for ensuring data quality, security, and compliance. It also involves mapping the relationships between different data sources and understanding how data flows through the organization.

Data lineage plays a vital role in this phase. By tracing the origin and transformations of data, you can gain a complete understanding of how data has been used and modified over time. This is essential for data governance, compliance, and impact analysis.

By following these five phases, organizations can gain a comprehensive understanding of their data landscape, laying the foundation for effective data management, analysis, and governance. For more details, a documentation workbook, and associated open-source tools refer to Appendix A.

Data Discovery Maturity Model

Resources should be applied to each phase as technology, business, and security requirements change. As such, the Data Discovery Maturity Models featured within this book are not necessarily sequential and can occur concurrently.



Chapter 3

Data Collection and Ingestion



Data Collection and Ingestion

Once data has been discovered and catalogued, data ingestion takes center stage. With an ever-expanding attack surface and the increasing volume of security data, efficient and reliable data ingestion is paramount. It is the foundation for effective threat detection, incident response, and proactive security measures. That being said, what is data ingestion?

What is Data Ingestion?

Data ingestion is the process of gathering raw data from various sources and transporting it to a destination where it can be analyzed, processed, and stored. This is the start of your data pipeline and like oil, your data needs to be collected before it can be refined. Once that is complete it will become the fuel needed to power valuable insights.

It is common for organizations to use a variety of security tools such as firewalls, intrusion detection systems (IDS), and antivirus software. Each of these tools generates logs filled with important security information. The data ingestion process then involves collecting those logs using agents, or by configuring the tools to send them to a central server. After all that is done, it then filters them by removing unnecessary information—like extraneous debug information.

Why it's Important

Having a mature data ingestion process is essential for several reasons. Security data needs to be collected and made ready for analysis as quickly as possible to enable real-time threat detection and response. A well-designed ingestion process can handle the increasing volumes and demand speeds of data without slowing down your systems. Last, it helps maintain data quality by incorporating checks and filtering mechanisms to ensure that only relevant and trustworthy data enters the data pipeline. All of this ensures that teams are optimizing the entire process of collecting and transporting data, reducing the burden on your source systems and minimizing storage costs.

How Data Ingestion Works

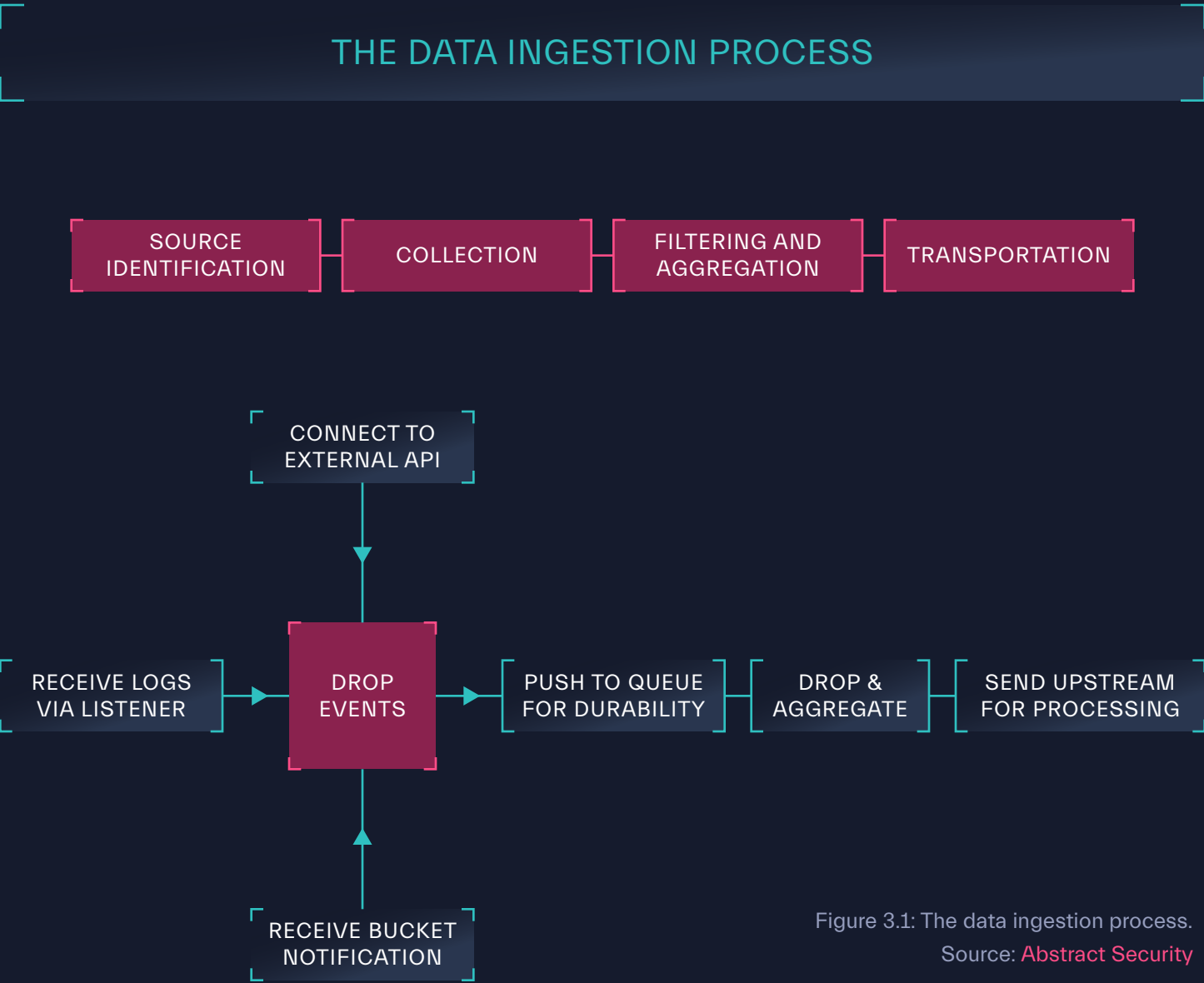


Figure 3.1: The data ingestion process.
Source: [Abstract Security](#)

The data ingestion process typically involves a few key steps (Figure 3.1). It starts with identifying all the relevant sources of security data, which could include logs from servers, endpoints, network devices, security tools, and cloud services. Next, teams need to collect this data from the identified sources. This can be done by either having the data “pushed” from the source to the ingestion pipeline—using methods like Syslog, HTTP listeners, or message brokers—or by having the ingestion pipeline “pull” the data from the source, using agents or API calls.

After the data is collected, it’s important to filter out any irrelevant information and aggregate similar events to reduce the overall volume and improve efficiency. Finally, the filtered and aggregated data is transported to its designated storage location, such as a data lake, data warehouse, or a SIEM system.

Reducing Unnecessary Data Costs

Reducing the processing of unneeded data is a key part of optimizing this task. For example, extraneous logging can be reduced by dropping health check patterns from trusted monitoring servers. Network traffic logs can be aggregated by specific criteria to reduce data volume while still capturing all the essential information.

Another useful tactic is to drop logs that require extra requests, such as permanent redirects from a web server, or similar HTTP upgrade requests. By carefully considering what data needs to be collected and how it will be used, teams can significantly reduce the overall cost of data ingestion.

Finding the Right Approach for You

Different industries have unique data ingestion needs. Financial institutions, for example, deal with highly sensitive customer data and must adhere to strict regulations. For this sector, organizations should carefully consider where to store data, how to ensure compliance, and what data needs to be readily available for analysis.

For other industries like Manufacturing, they may need to rely on real-time data from production lines, sensors, and supply chain systems to optimize operations and maintain quality control.

Optimizing Your Own Data Ingestion Processes

To validate that data sources are being ingested as intended, we have included a log onboarding checklist ([Appendix D](#)). Using this resource, uniformity can be ensured among different data feeds.

The checklist included provides helpful resources to aid in reducing data. This will enable creation and implementation of event filtering playbooks.

Playbook	Data Feed	Objective	Filter Applied
Keep-Alive & Health Checks	Various	Reduce extraneous logging	Drop health check pattern originating from trusted health monitoring servers
Network Log Aggregation	Various Network Devices	Aggregate logs while enabling full signaling	Aggregate network traffic by Proto Src SrcPort Dest DestPort User and count RX + TX bytes
Drop Permanent Redirects	Load Balancer Web Server	Reduce logs which require extra requests	Drop HTTP status codes 301, 304, 308, 1xx
Drop No Value Events	Various	Prune logs with no operational use cases	AWS: Successful EKS Encrypt-Decrypt Successful EKS GetInstanceProfile Successful GenerateDataKey events
Drop Service Account Traffic	Various	Drop all normal service account operation	Drop successful requests for service account traffic from known monitoring/logging systems

Data Ingestion Maturity Model



Chapter 4

Data Processing



Data Processing

After successfully ingesting and storing security data, the next critical stage in a data strategy is data processing. This is where raw data is transformed and refined into a usable format.

What is Data Processing?

Data processing involves a series of operations that cleanse, transform, and organize data to make it suitable for analysis, reporting, and decision-making. In this cycle, data can undergo many enrichment, normalization, distillation, and filtering steps towards its end goal of storage (Figure 4.1).

This can include:

- DATA CLEANSING**
Identifying and correcting errors, inconsistencies, and missing values in the data.
- DATA TRANSFORMATION**
Converting data from one format to another, aggregating data, and enriching it with additional information.
- DATA REDUCTION**
Reducing the volume of data by filtering out irrelevant information or aggregating similar events.
- DATA DE-IDENTIFICATION**
Removing or adjusting sensitive data to meet regulatory obligations for privacy protection and reduce risk.

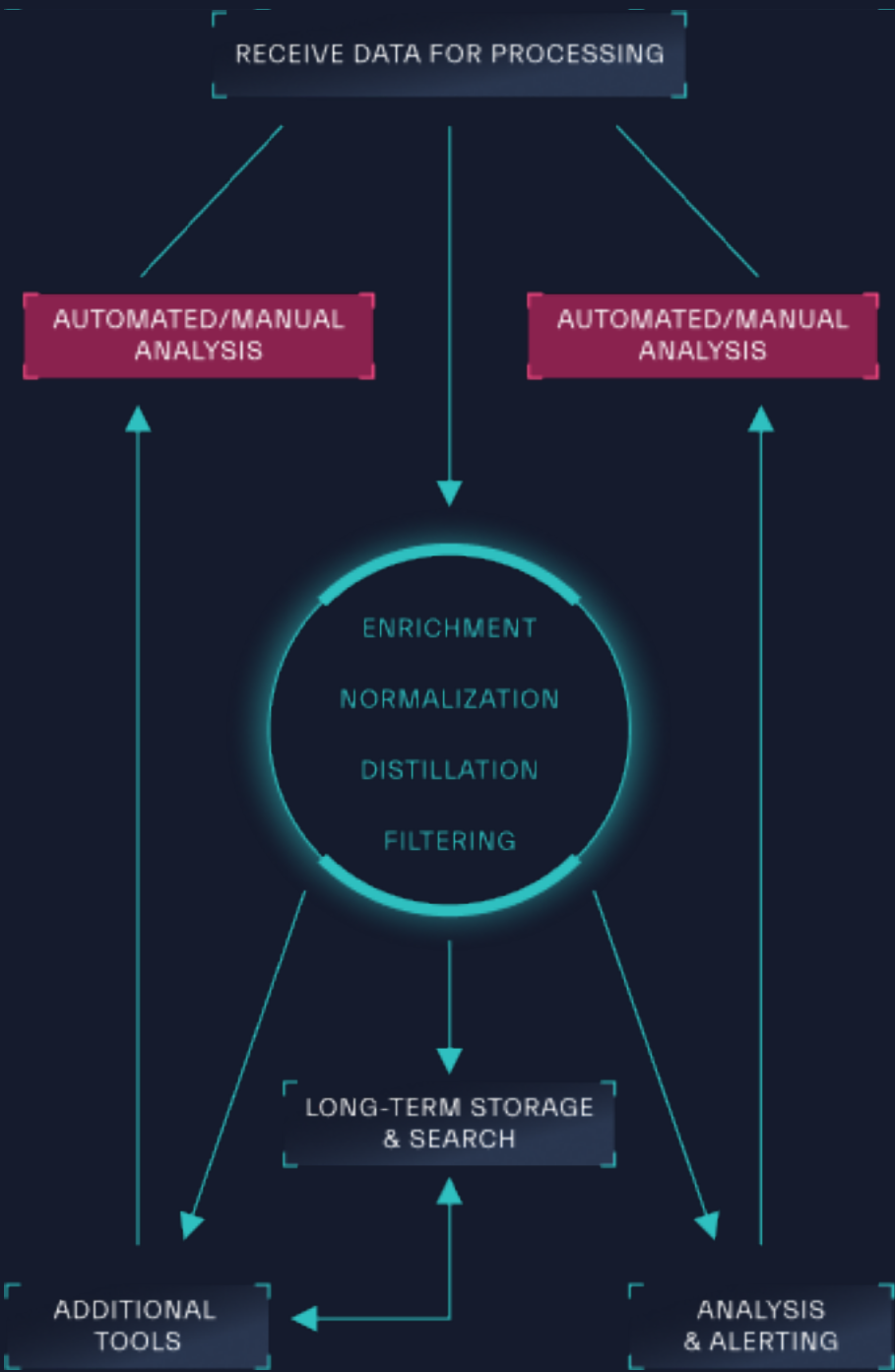
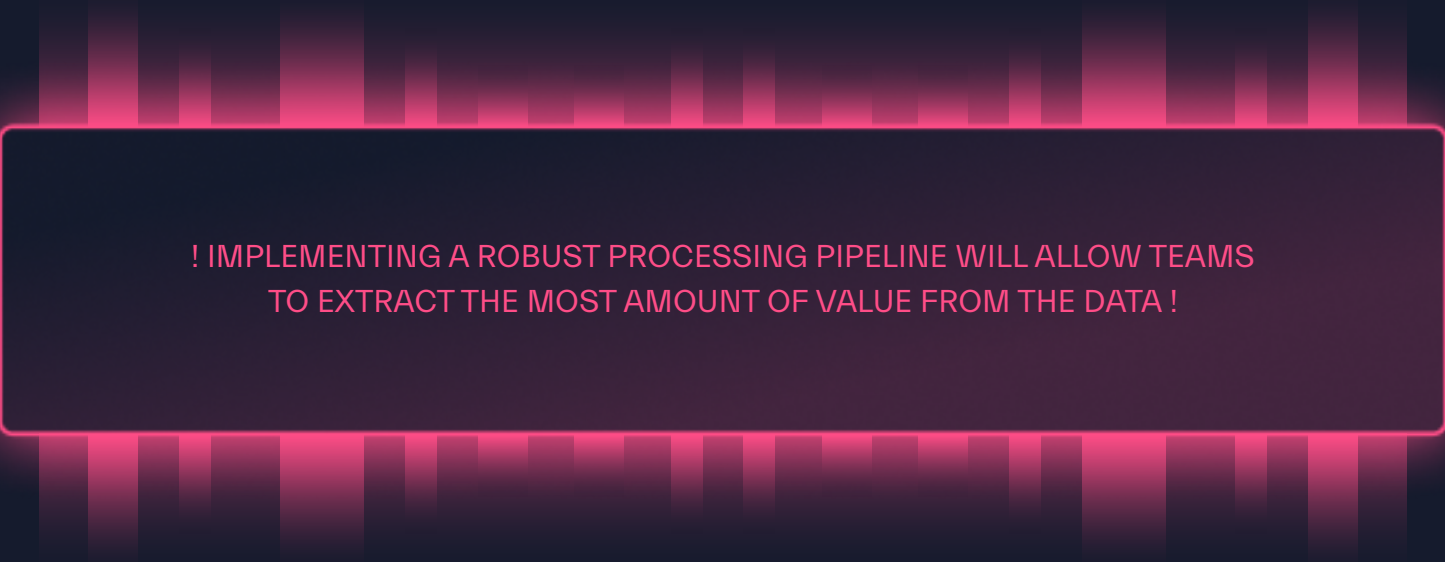


Figure 4.1 Data processing. Source: Abstract Security

Data Normalization: From Raw Data to Refined Insights



Compliance and operations requirements should dictate what data needs to be processed. Having a data normalization guide will help ensure consistency when onboarding new data sources:

Playbook	Description
Timestamp standardization	Identify all time stamps and corresponding formats and define a primary timestamp field.
Time zone normalization	! Wherever possible align timestamps with UTC ! for consistency and uniformity across sources.
Fields with use cases identified	Define which fields provide data with existing use cases. Note fields which would be candidates for tuning in ingestion phase.
Text normalization	Clean up field data by removing leading and trailing spaces, duplicating fields, casting to lowercase, and sanitizing input if applicable.
Field mapping to schema	Map fields to common schema such as Elastic Common Schema (ECS), Splunk Common Information Model (CIM), Common Event Format (CEF) or Log Event Extended Format (LEEF).
Sensitive information validation	This should have already been performed in discovery; however, re-validating will ensure early detection should any mistakes have been made and de-identify or redact any data that should be removed.

Checks and balances should be enacted to continuously validate data quality and avoid structure drift. This happens when data undergoes slight changes in the underlying data schema or format which can potentially invalidate models trained on the older data structure. Having a monitoring and testing checklist can also help evaluate data sources involved in the processing phase.

Monitoring and Testing Checklist

Checklist Item	Example Tests
Health Monitoring	• Up/down alerting for data source • Error rate threshold alerting • Data spike and drop alerting
Data Integrity Monitoring	• Field removal or change alerting • Duplicate field detection • Missing field or value change to only empty
Performance Monitoring	• Latency threshold exceeded • Resource utilization alerting • Queue size increases
Security Monitoring	• Sensitive data discovery • Access log & audit validation • Encryption & Data Loss Prevention (DLP) validation
Disaster & Resiliency Testing	• Failure & retry logic validation • Failover testing • Backup & restore testing

Evolving Data Processing Architecture

Historically, many data processing architectures involved a single destination for processed data. However, as data volumes and business needs have evolved, best practices now advocate for a more flexible approach that allows for multiple consumers of processed data through data bifurcation, splitting the data targets into multiple branches or repositories. A publisher/subscriber architecture like this allows for greater scalability and adaptability.

Key Data Processing Techniques

Governance and Locality

Consider where data should be processed and stored, especially due to regulations like GDPR and FedRAMP. A distributed architecture may be necessary to comply with data residency requirements and enable global operations.

To scale operations globally, architectures must be adopted to be able to process and store substantial amounts of data within the boundaries of the business; all while decoupling alerts and notifications.

Enrichment

Data enrichment can take various forms. Some examples are creating metadata topics, tables, and indexes for additional models to help inline functions that add context to processed data. These kinds of enrichments can help increase the value received from data sources, while also enabling new functions that would have otherwise been impossible.

Scoping enrichment usage and volumes will allow appropriate resource sizing. Some enrichments can happen quickly using local data, while others may require additional time to poll an external data source such as an API or read a remote file.

Understanding the number of events per second (EPS) to support, as well as the CPU and memory requirements surrounding enrichment, will allow resource allocation as required. This will also provide a benchmark to help with future enrichment resource forecasting. All enrichments should be mapped to use cases so that complexity and storage costs aren't needlessly increased.

Aggregation and Summarization

Depending on current tools, there may be opportunities for aggregation during the data ingestion and processing phases. The key goal is to reduce the amount of data while not losing any signaling capability.

A wide variety of sources can be summarized to reduce the amount of data required to perform effective signaling and analytics. For example, domain name system (DNS) logs within an environment can be quite large in volume. Creating batch jobs, which summarize a full day’s worth of questions and corresponding answers, may take a considerable time to run. However, in most cases this is due to there being many repeated values in the data.

If this data can be counted and sources stripped out, organizations could significantly reduce the amount of data being summarized. 20 gigabytes (GB) of DNS logs could potentially turn into 100 megabytes (MB) through a daily aggregation process, drastically reducing the cost of storage. This same methodology can be applied to dozens of other categories of data.

Saving Time and Resources

If a piece of data is found which requires further investigation, it must be searched for and any corresponding data returned. Using summarized indices allows crafting a much faster secondary query, targeting only the time periods which contain the data. For example, instead of searching through one year of data, only the last 24 hours of un-summarized data would have to be searched, as well as the summarized data, and any associated raw data.

Type of Records	Type of Records	Type of Records
DNS	Daily count of question/answer pairs	• New domain detection • First seen enrichment • Traffic spike from new domain
Proxy	Daily count of user/method/domain tuple	• File download from new domain • “Safety in Numbers” (Appendix G) • Data sent to new domain
Authentication	Daily count of source/user/target/outcome tuple	• Login to new system • Login from new source
Email	Daily count of sender/recipient/domains in links/attachment extensions	• Suspicious link from new sender • Suspicious attachment from new sender

Normalization

Standardize data formats and structures to ensure consistency and facilitate analysis. It also enables a common taxonomy when dealing with data and logs. This can be leveraged very effectively with the adoption of a standardized field dictionary and a standardized language to describe queries, like Sigma (Appendix H).

Filtering

With the goal to increase signal and reduce noise, opportunities to reduce fields or data may arise. This is especially useful for cloud-based components which can accrue costs by the minute. The data reduction and aggregation playbooks included earlier will provide an excellent starting point for filtering ideas which may also be applied to processed data before it is sent to its destination.

Distillation

Distillation is the act of extracting meaning from logs by identifying the most actionable and relevant information. While distillation can be considered part of the normalization portion of processing, it is worthwhile to take time to understand exactly why the data is being analyzed, what makes it actionable, and what is relevant to decision-making and generating insights.

Flight Recorder

Flight recorder data can be voluminous and transient, whereas authentication logging or proxy connections have longer, more useful lifespans. Where possible, it is advantageous to have an unfiltered view of all your data, even if only for a few minutes or hours. While storage costs to keep data long term might be astronomical, it is often palatable to dedicate a fixed amount of storage and compute as an unfiltered rolling buffer of data. This is useful in operational and security incident work, as well as performance troubleshooting.

This model can also enable short term visibility for enhanced review of highly detailed log sources. Distilled data can meet regulatory log retention obligations that many organizations face without drastic cost increases.

Feedback Loops

Iterating when dealing with separating signal from noise is crucial. In the cat and mouse security world, a good countermeasure can cause an adversary evolution. Building deliberate data pipelines which allow defenders to evolve analysis and alerting faster than attackers is how the costs to the adversary are raised. In some situations, it may be more profitable for an adversary to move on to a softer target which evolves more slowly.

Data Processing Maturity Model



Chapter 5

Data Storage



Data Storage

What is Data Storage?

Data storage, in the context of security, is far more than just finding a place to write logs. It’s about strategically organizing and managing security data to ensure it remains accessible, secure, and optimized for analysis. Think of it as building a fortress for valuable intelligence. The design needs to be robust, adaptable, scalable, responsive, and able to withstand the test of time (and potential attacks).

Depending on an organization’s maturity, storage implementations can vary significantly. Less mature organizations might rely on a single platform like Google Cloud Storage or an Amazon S3 bucket. Some organizations may split the storage for cost-effectiveness or speed of access. However, as data volumes and processing needs grow, a more sophisticated approach becomes necessary, often involving a data lifecycle management process with hot, warm, and cold tiers to balance cost-effectiveness with access and latency requirements (Figure 5.1).

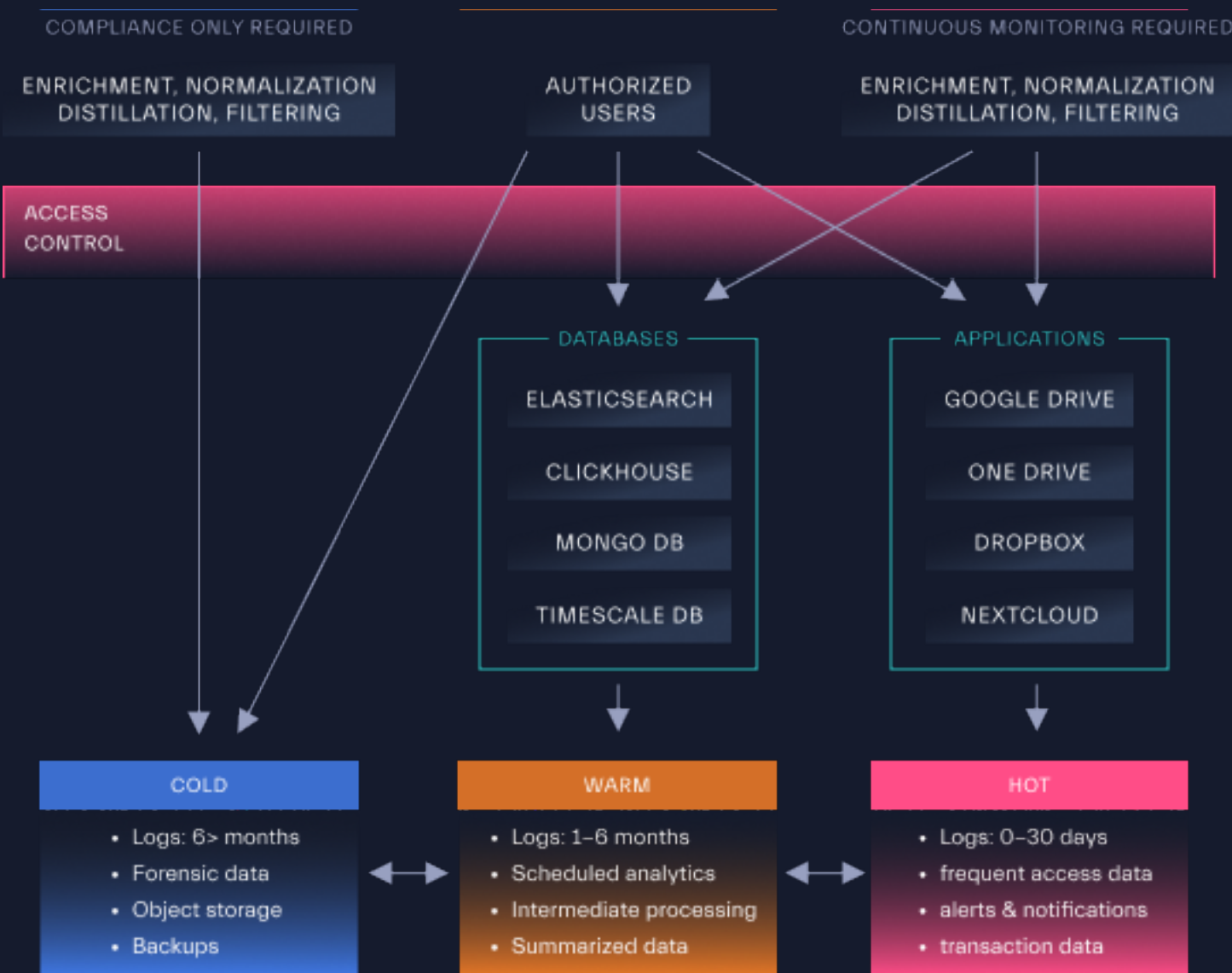


Figure 5.1: Access Control Based on Hot, Warm, and Cold Storage Requirements. Source: Abstract Security

Regardless of the organization’s size, establishing standards is imperative. Implementing naming standards across the environment can significantly reduce search time, reduce costs, prevent duplication, and streamline the validation of backup policies. Another often-overlooked element, especially in smaller environments, is a regularly updated and tested backup and retention policy ([Appendix J](#)). This policy is a safety net, enabling rapid recovery in the event of a catastrophic data loss incident.

Data Storage Tiers and Search Requirements

Business requirements and resource costs often necessitate a tiered approach to data lifecycle management. Consider the following examples and their respective storage needs based on search and retrieval requirements:

Application and System Transaction Data	This data demands millisecond retrieval for real-time usage and is best suited for hot storage solutions like relational databases (PostgreSQL, MariaDB), Redis, or MongoDB.
Recent Security Logs (0–30 Days)	Frequently accessed for alert triage and incident response, these logs also require hot storage solutions like Elasticsearch, Clickhouse, or Splunk.
Real-time Analytics Data	Continuously accessed for dashboards and real-time monitoring, this data necessitates hot storage with tools like Elasticsearch or Clickhouse.
Aggregated Data and Scheduled Analytics	Accessed with medium frequency for summarized reports, this data can reside in warm storage, utilizing Elasticsearch, Clickhouse, or relational databases.
Historical Logs (1–6 Months)	Periodically accessed for investigations or audits, these logs are well-suited for warm storage with Elasticsearch or Clickhouse.
Backups	Rarely accessed except for testing or disaster recovery, backups are best stored in cold storage using object storage, Veeam, or other managed backup solutions. Backups are best kept offline whenever possible.
Archived Logs (6+ Months)	Accessed infrequently for compliance or forensic investigations, these logs should be stored in cold storage, leveraging object storage, NAS, or SAN solutions.

By clearly defining retrieval requirements, organizations can evaluate the tools and technologies that best align with their business needs while adhering to operational and regulatory obligations. Features like Write Once Ready Many (WORM) or other immutable storage solutions should also be considered to prevent data modification or tampering, especially when compliance mandates data integrity.

Defining Hot, Warm, and Cold Storage Tiers

Storage performance tiers must be defined clearly to ensure the data that resides there is responsive to the needs of the organization for each data type. Vendors will often prescribe connectivity requirements but a clear understanding of your organizations' needs and the data that you are ingesting, processing, and analyzing can allow you to adjust your storage architecture to reduce costs but easily meet or exceed operational requirements.

Things that affect performance include Input/Output Operations Per Second (IOPS), Redundant Array of Inexpensive Disks (RAID) type, bandwidth, throughput, latency, write wear cycle, connectivity type, and connectivity protocol.

IOPS are more important than any other value for high-speed data connections or millisecond response. Consider flash or other solid state drive technologies.

RAID type has a high impact on high write operations. The performance impact of all RAID types reduce the possible drive IOPS. RAID1+0 is appropriate for most log targets, but other RAID types (RAID6, RAID50) may be a feasible option for archive or slow targets while reducing costs substantially.

Bandwidth and throughput are important to be aware of to meet your data access needs, but IOPS are typically a higher value metric to pay attention to. Bandwidth is a factor of the medium that data is transported over. throughput is a factor of the bandwidth multiplied by the possible IOPS. On a network, when considering NFS, SMB, or object storage connections, network considerations can be as important as volume considerations.

Latency is the delay in time to respond to a read or write request. This is typically measured in milliseconds and can lead to delays in log writes, analysis activities, or populating queries or dashboards. Latency can be caused by a number of things but the most likely is due to the architecture of the solution and physical nature of the devices or network between the requester and the read or write response.

There will always be a performance bottleneck, what you need to determine is how much money you have available to spend, and where you want to place that bottleneck for the least impact given your budget.

Hot Tier	High bandwidth, high throughput, high write operation, low latency. Typically, local disk using flash or SSD drives with a high IOPS or fiber channel Storage Area Network (SAN) storage using RAID1+0 is appropriate.
Warm Tier	High-mid bandwidth, high-mid throughput, mid or scheduled write operations, low to mid latency. RAID1+0 or RAID5 disk using flash, or traditional disk is possible connected via fiber channel SAN, Network File System (NFS), or Amazon S3 storage where vendor supported.
Cool Tier	Also known as the archive tier. Low bandwidth, low throughput, mid to high latency, low write, low read operation. RAID5 or RAID6 using traditional disk connected via NFS, Server Message Block (SMB), or Amazon S3 storage where vendor supported.

Streamlining the Data Lifecycle

Once basic standards and policies are established, organizations can focus on streamlining their data lifecycle process. Data locality is a key consideration, especially given regulations like GDPR and FedRAMP, which might mandate data processing and storage within specific regions. Adjusting the architecture to ensure compliance with these requirements is crucial for seamless global operations.

A well-defined data storage strategy must also consider the diverse needs of various teams within an organization's data ecosystem. Clearly articulating data types, tolerable latency, query complexity, and data visibility requirements will help determine the most suitable tools for each team. Evaluating these needs holistically can lead to tool reuse and ensure that each solution is fit for its purpose.

For example, SOCs require real-time access to raw data for incident analysis, while Hunt and DFIR teams might need both real-time and batch access for investigations. Compliance teams may have ad hoc access needs with specific data visibility requirements, and legal teams might require access to raw data for legal proceedings.

In some cases, retaining a limited buffer (1–48 hours) of completely raw, unfettered logs can be invaluable for incident response or root cause analysis related to performance issues. This increased visibility, while potentially resource-intensive, can significantly expedite troubleshooting, persist post-event artifacts discovered while investigating potential false-negatives, and provide valuable feedback for optimizing data processing pipelines.

As organizations grow, streamlining data lifecycle, ingestion, and access processes becomes increasingly critical. Bottlenecks that might not be apparent at smaller scales can emerge, and tradeoffs often exist between ingestion time and query time. Clearly defining use cases and latency tolerances will guide the necessary tuning for the storage environment.



Data Storage Maturity Model



Chapter 6

Data Analysis



Data Analysis

What is Data Analysis?

Data analysis is the engine that drives security data strategy. It's the process of transforming raw data into actionable insights, enabling teams to understand, respond to, and even predict security threats. Think of it as the art and science of deciphering the clues hidden within the data, revealing patterns, anomalies, and ultimately, the knowledge necessary to protect the organization.



The Data Analysis Process

Analysis helps break down complex security challenges into manageable components, allowing teams to extract meaningful insights. To help jumpstart analytical capabilities, check Appendix K for real-world use cases from various organizational functions, including finance, SRE, and security. These use cases can be adapted, implemented, or serve as inspiration for developing analytics initiatives.

Effective data analysis is an iterative process that involves several key stages (Figure 6.1):

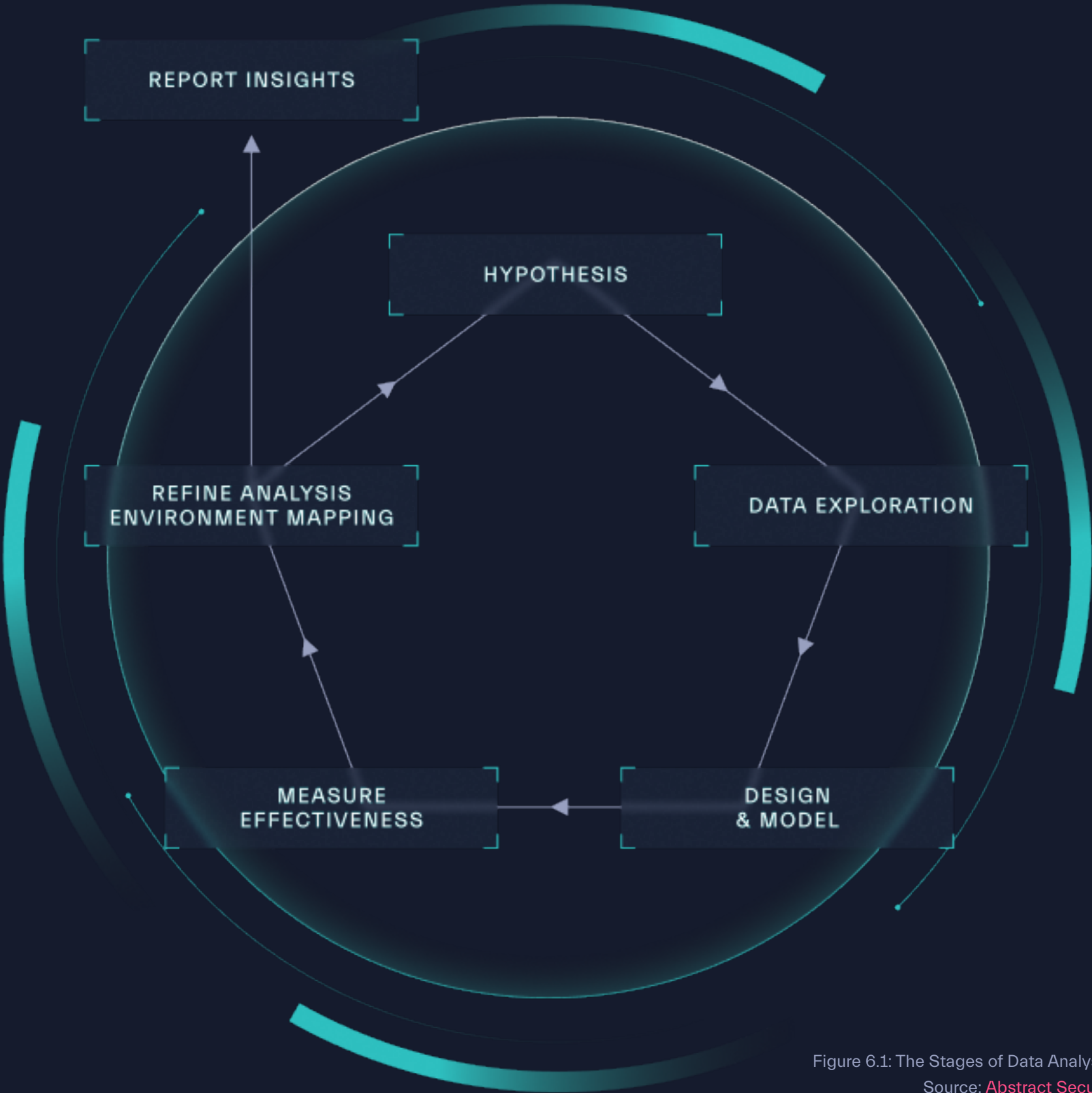


Figure 6.1: The Stages of Data Analysis.
Source: [Abstract Security](#)

STAGE 1

Business-Driven Input and Hypotheses

Analysis begins with defining a problem to be solved. Reporting requirements play a crucial role in defining what needs to be analyzed, how soon, and why. Given the constant influx of security challenges and limited resources, prioritizing these problems is essential. Factors like criticality, risk, impact, time sensitivity, and strategic alignment should guide prioritization efforts.

STAGE 2

Exploratory Data Analysis (EDA)

Data profiling is crucial to assess whether the available data supports the proposed idea. Analyzing trends, outliers, and distributions can reveal potential correlations between data attributes and desired outcomes. Identifying common and rare values within the data helps refine the scope of the analysis and uncover potential insights.

STAGE 3

Relationship Modeling

Entity Relationship Diagrams (ERDs) are valuable tools for visualizing the connections and relationships between different data components. These models can illuminate potential issues within business processes and uncover hidden information that can improve analysis results. Graph-based techniques, utilizing graph databases and algorithms such as ‘shortest path’, can further enhance the exploration of complex relationships and dependencies. Statistical measures, such as Chi-square tests and correlation coefficients, can be used to evaluate the strength and significance of these relationships.

STAGE 4

Pattern Analysis

Identifying patterns within security data is crucial for uncovering anomalies and potential threats. This involves searching for recurring sequences, trends, and outliers that might indicate malicious activity. Techniques like time series analysis, clustering, and anomaly detection algorithms can be employed to uncover these hidden patterns.

Data visualization plays a vital role in pattern analysis. Heatmaps, for example, allow multiple data attributes to be visualized simultaneously, revealing correlations and outliers that might otherwise go unnoticed. Analyzing the frequency and distribution of events over time can provide valuable insights into the behavior of users, systems, and applications.

STAGE 5

Designing and Evaluating Models

Analysis models can take various forms, from simple statistical analysis to complex machine learning algorithms. Clustering, regression testing, feature engineering, and dimensionality reduction are valuable techniques to consider when designing models. Before diving deep into model design, it's crucial to clarify stakeholder objectives and ensure the chosen model aligns with the problem the organization is trying to solve.

Prioritize simple and explainable models whenever possible, especially in the early stages of development. Plan for the models to evolve and be easily updated as new data becomes available or requirements change. Discrepancies in data, ingestion issues, or processing errors might necessitate adjustments to the data preparation and model design. Continuous exploration, refinement, and validation are key to developing effective models.

STAGE 6

Measure Effectiveness

How well does the model perform? What percentage of the time is it effective? Asking and answering questions like these will guide refinement efforts, increasing model effectiveness and ultimately helping to achieve analysis objectives. Additionally, determining whether extra steps should be taken to increase effectiveness is recommended. Regression testing can also further validate findings and provide actionable recommendations.

STAGE 7

Autonomic Responses

In some cases, highly effective models can be implemented through autonomic processes. For example, Security Operations can benefit from models that automatically identify and block traffic from malicious sources based on predefined criteria. These models, once rigorously tested, can autonomously respond to threats, freeing up valuable human resources.

STAGE 8

Refining and Reporting Insights

Continuous refinement is essential for maintaining the effectiveness of analysis models. Acknowledge potential biases in data or analysis methods and ensure models can scale to meet future demands. Clearly document and communicate the findings, ensuring the insights derived from analysis are readily accessible to stakeholders.

The goal is to develop models that can be seamlessly integrated into Security Orchestration, Automation, and Response (SOAR) platforms, providing actionable insights with minimal human intervention. These high-fidelity models can automate tasks, improve threat detection, and optimize security processes.

Data Analysis in the Real World

Different industries face unique data analysis challenges:

Financial Services

Financial institutions often leverage real-time analytics to detect fraud, monitor transactions, and ensure compliance with regulations.

Healthcare

Healthcare organizations analyze patient data to improve care, identify trends, reduce patient risks, enhance safety, and enhance operational efficiency.

Manufacturing

Manufacturers analyze data from production lines and supply chains to optimize processes, predict maintenance needs, and improve quality control.

Data Analysis Maturity Model



Chapter 7

Data Reporting



Data Reporting

Discovering, collecting, processing, and analyzing data is only valuable if its insights are being effectively communicated to the right people at the right time. Reporting is the bridge that connects data-driven discoveries with decision-makers who can drive change within an organization. Effective data reporting is all about presenting information concisely so that stakeholders can make informed decisions about what choices they should, or should not make, and why (Figure 7.1).



Figure 7.1:
The Stages of Data Reporting.
Source: [Abstract Security](#)

Depending on the business unit, stakeholders may need to be provided with various types of reports per their specific needs (Figure 7.2):

Report Name	Purpose	Frequency	Business Unit	Example Insights
Executive Summary	High-level overview of organization performance	Weekly / Monthly	Executive Leadership	Revenue, Profit, Margins, KPI's
Sales Performance	Track sales team achievements and opportunities	Weekly	Sales	Conversion Rates, deal pipeline, top clients
Marketing ROI	Measure effectiveness of marketing campaigns	Monthly / Quarterly	Marketing	Campaign metrics, lead generation, ROI
Financial Forecasting	Projections for budgeting and resource allocation	Monthly / Quarterly	Finance	Revenue projections and expense trends
Compliance Reporting	Provides evidence for compliance to regulatory frameworks	On Demand	Risk + Compliance	SOC2, GDPR, ISO, 27001, and other regulatory frameworks
Security Incident reporting	Summarizes threats, incidents, affected resources and resolutions	Daily / Weekly	Security Operations	Threats detected, patching coverage, mean-time-to-resolution
IT Infrastructure health	Monitor status of network and systems operations	Daily / Weekly	IT	Uptime, bandwidth usage, performance metrics, incident trends
Product Development metrics	Tracks product roadmap progress and user engagement	Monthly	Product Management	Feature adoption, development velocity

Figure 7.2: Common Types of Reports. Source: [Abstract Security](#)

The Data Reporting Process

Effective data reporting involves several key steps:

1. Define Objectives

Clearly articulate the purpose of each report and the specific information it should convey. Consider the target audience, their needs, and how they should be using the report. For example, operations teams may need more continual insight into platform metrics, while Finance will need more information regarding licensing costs and forecasting data.

Taking the time to carefully craft detailed objectives will lead to more effective scoping and a faster achieved outcome. Early introduction of caveats can also help clear any likely misconceptions.

1. Design Reports

Create visually appealing and informative reports that are easy to understand and navigate. Use appropriate visualizations, such as charts, graphs, and tables to present data effectively. Be as direct as possible and make sure there isn't any ambiguity involving the data that is being delivered.

1. Automate and Deliver

Automate report generation and delivery to ensure timely and consistent reporting. Consider using scheduling tools or dashboards to provide real-time access to key metrics. Additionally, automation will be useful to discover outliers and data discrepancies.

1. Stakeholder Review

Gather feedback from stakeholders to ensure the reports meet their needs and expectations. Continuously refine and improve reports based on user feedback. Measure delivery, usage, and most importantly, whether they are reaching their desired outcomes using the reports. Engagement is equally important to extracting the right insights and iterating through adaptations.

Data Reporting Maturity Model



Chapter 8

Data Governance & Security



Data Governance & Security

A well-implemented governance process is a powerful business enabler, providing a structured framework for innovation that allows organizations to safely and effectively leverage their data assets. Transparency and scalability can be enhanced by implementing a centralized access matrix document or database, as seen in [Appendix B](#).



What is Data Governance?

Policies and standards provide further structure, clarifying regulatory requirements and guiding data usage. Predefined policies and standards empower teams to develop tailored procedures for leveraging data within their specific environments. This helps foster innovation and accelerates the adoption of new tools and technologies, enabling organizations to achieve their strategic objectives (Figure 8.1).

However, striking the right balance between oversight and flexibility is crucial. Too little governance can expose you to legal and financial risks, while excessive governance can stifle growth and productivity. A successful governance program is one that effectively mitigates risks without hindering innovation.

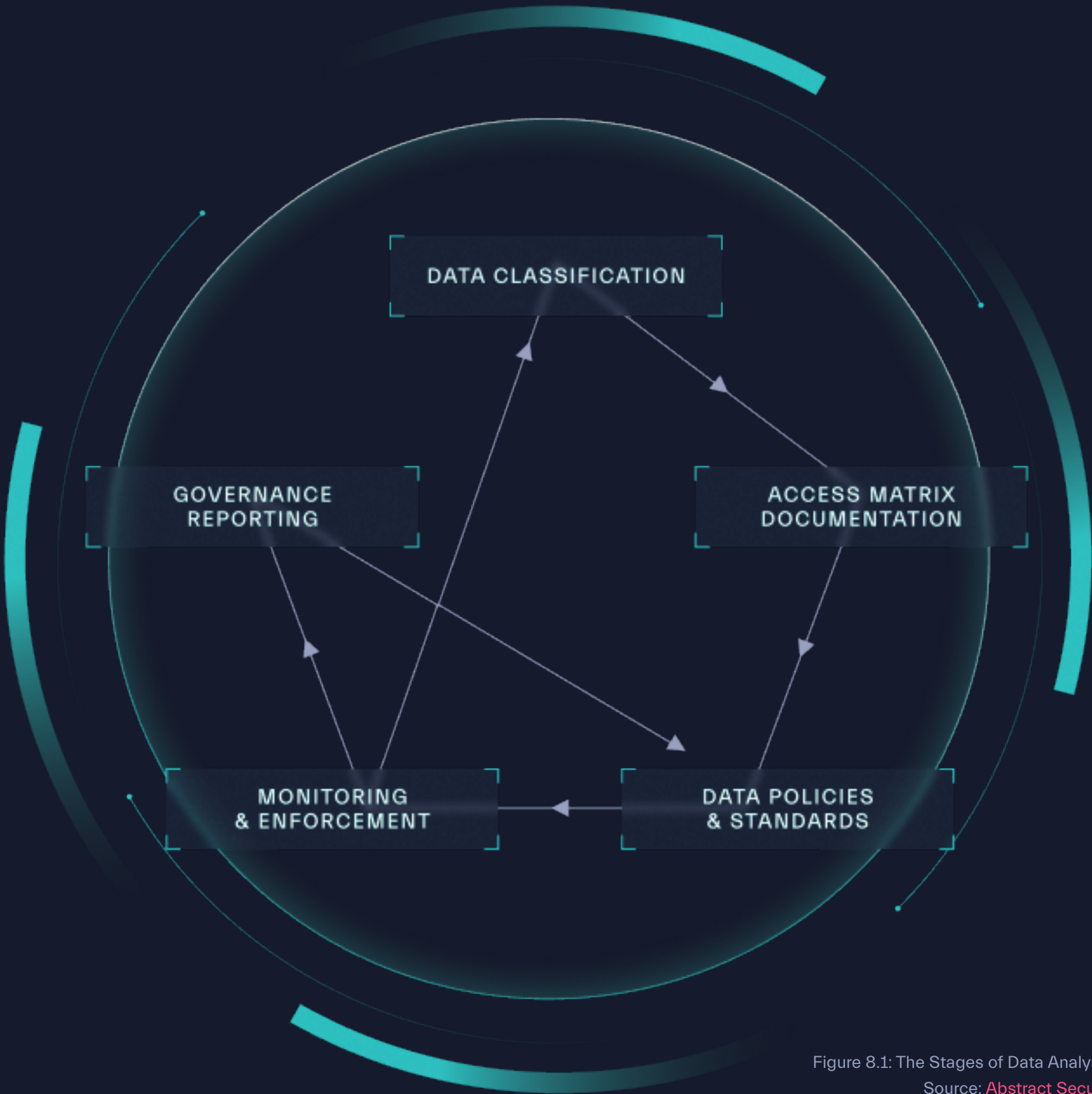


Figure 8.1: The Stages of Data Analysis.
Source: [Abstract Security](#)

Key Components of Data Governance & Security

1. Data Classification

Data sources need to be classified within the asset inventory and access matrix. This ensures compliance with relevant laws, regulations, and industry standards. Factors like data sensitivity, regulatory requirements (GDPR, HIPAA, PCI and others) and internal policies all contribute to the classification process.

Special care must be taken when dealing with sensitive data in conjunction with mass indexing systems, machine learning algorithms, and Large Language Models (LLMs). Implementing access controls and supporting documents, as dictated by applicable policies and standards, allows for effective monitoring, enforcement, and reporting.

“Data classification is so critical because it lies at the heart of setting data privacy, protection, recovery, and governance requirements. In addition to facilitating compliance, this is necessary to ensure that all data is recovered post-breach and to prioritize recovery so that the organization’s “crown jewel” data assets are recovered first.”

Krista Case, Research Director, Futurum Group

1. Access Matrix Documentation

A well-maintained access matrix is essential for scaling data operations effectively. This centralized document will serve as a single source of truth for data ownership, permissions, and applicable policies and regulations. It should be closely tied to the Identity and Access Management (IAM) and provisioning processes, dictating who has access to what data, when, and for how long.

Automation and scripting can help streamline the collection of access-related information from disparate systems. Creating a centralized IAM database to track users, groups, roles, and permissions provides a valuable resource for various purposes, such as DFIR investigations, resource usage analysis, license planning, and audit support.

1. Data Policies and Standards

Ultimately, the data policies and standards adopted need to align with the organization's compliance requirements and business goals. They should be driven by industry best practices and regulatory mandates, providing clear guidance on data handling, access, retention, and security. For a starting point on developing such data policies, check out [Appendix C](#).

Clarity and accessibility are key. Policies and standards should be available to all staff and written in plain language so that it is easily understood by all stakeholders. Avoid technical jargon. Clearly defined scopes and links to related processes and procedures further enhance the effectiveness of a governance program.

Staff must feel empowered to implement and adhere to defined policies and standards even when challenged by a superior and especially when urgency is applied. Policies must be reviewed on a regular periodic basis or at least bi-annually to align with changes to governance frameworks, regulations, and laws.

1. Monitoring and Enforcement

Governance and Security Operations teams share the responsibility for monitoring and enforcing data policies. Documenting escalation paths for data-related incidents and clearly defining data ownership within the asset inventory can significantly reduce response times and ensure accountability. A well-defined exception process can reduce the impact of unexpected changes and events, and provide organizations with flexibility while still applying governance practices.

Data health and continuity requirements should also be aligned with business needs. Wherever possible, data should be pruned and reduced to minimize storage costs and processing overhead. Proactive deprovisioning of unused accounts and data masking or de-identification techniques can further enhance security and reduce risk.

1. Governance Reporting

A well-maintained asset inventory and access matrix enable the generation of comprehensive governance reports. These reports provide data owners with insights into their security posture, highlight potential risks, and facilitate proactive remediation efforts. They can also be used to support audits, inform business growth initiatives, and prioritize resource allocation.

“Attack surfaces are sprawling due to the rise of multi-hybrid cloud infrastructures and the booming ecosystem of SaaS applications and IoT devices. At the same time, controlling machine identities is only becoming more challenging as the usage of agentic AI increases. With this in mind, it is no surprise that identity-based attacks are becoming favorites of malicious actors – and it follows that there is arguably nothing more critical to data security and privacy than modernizing data access controls. Granular, just-in-time, and conditional access to data has become paramount.”

Krista Case, Research Director, Futurum Group

Data Governance Maturity Model



Chapter 9

Understanding Common Data Platforms & Tools



Understanding Common Data Platforms and Tools

With an understanding of what an effective security data strategy looks like, it is time to determine what kind of platform and data tools are right for your organization. From talking with security practitioners, we realized that most enterprise organizations have at least four different repositories for security data, including SIEMs, XDR/NDR/EDR tools, vulnerability scanners, data lakes, and log management systems.

These repositories store data from a multitude of different applications, sensors, security devices, and applications. The marketplace is inundated with acronyms, and it can be confusing for organizations and business leaders who are unfamiliar with the diverse set of data platforms and where the return on investment is best placed across the alphabet soup.

DISCLAIMER: PLEASE NOTE THAT THE VENDORS PROVIDED LATER ON IN THIS CHAPTER ARE FOR EXAMPLE PURPOSES ONLY, AND WE DO NOT OFFICIALLY ENDORSE, SUPPORT, OR SPONSOR THEM IN ANY CAPACITY.

What is a SIEM?

Security Information and Event Management (SIEM) systems ingest logs from a multitude of sources across an organization’s IT environment including servers, firewalls, intrusion detection systems, and much more. Well deployed and managed SIEMs then correlate these events to identify patterns, anomalies, and potential security threats. They provide a birds-eye view of the organization’s entire security posture, helping to consolidate its data in one place, with the opportunity to delve deep if necessary.

When addressing security operations, most large organizations turn to SIEM systems. This has been the case since the early 2000s when pioneering vendors like ArcSight, eSecurity, Intellitactics, and Net Forensics, offered first-generation SIEM systems. While SIEMs are still in use, security professionals typically face multiple SIEM challenges (Figure 9.1):

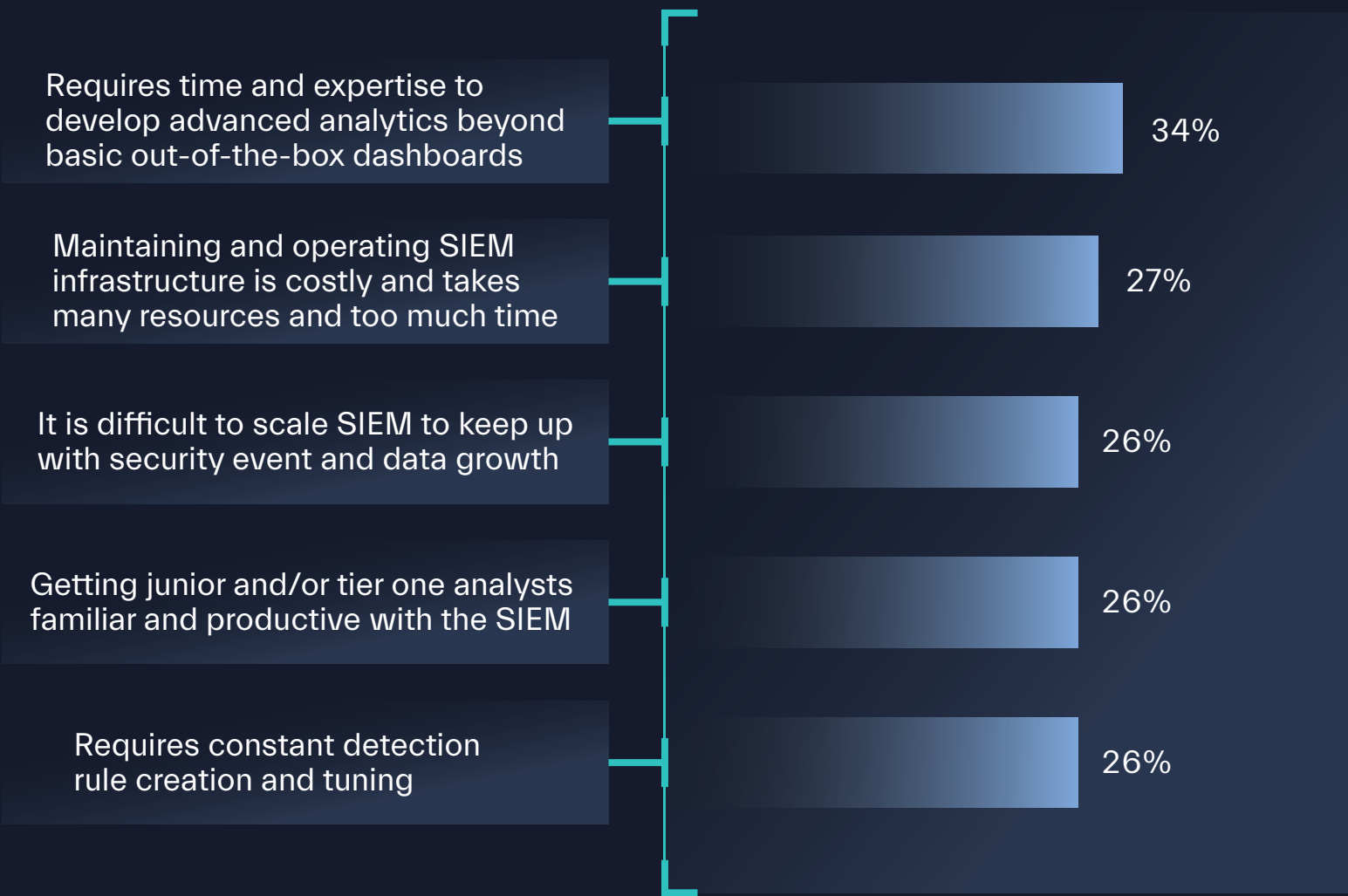


Figure 9.1: SIEM Survey. Source: [Enterprise Strategy Group](#)

Security operations professionals often complain that SIEMs are costly to buy/operate and lack workflow functionality. Tactical adjustments to security operations technologies and processes struggle to keep up with the dynamic changes and growing scaling needs of security operations.

SIEM Pros and Cons

PROS	CONS
Offer real-time analysis and alerting capabilities, helping security teams to detect and respond to incidents quickly.	SIEMs can be difficult to successfully implement into existing IT infrastructure. Lots of time and resources will need to be dedicated to solving compatibility and data format issues.
Many solutions offer advanced analytics, cloud integration, and user and entity behavior analytics to enhance threat detection capabilities.	Can be very hard to operate and may require extensive training or additional specialized hiring. Additionally, it may also require relying on an outside vendor to operate and manage, which can complicate its maintenance and scalability.
Can streamline forensic investigations by providing detailed and actionable logs.	SIEMs tend to have a high rate of false positives, which require continuous tuning and deployment to keep from overwhelming security teams.

Sample Vendors to Research

- Splunk Enterprise Security
- IBM QRadar
- Blumira, SumoLogic, Exabeam, Logrhythm, Sentinel, Securonix, Chronicle, Elastic
- Logz.io, NetWitness, Odyssey, QAX, ManageEngine, Logpoint, Devo, Grucul, Rapid7, Fortinet
- Graylog, SIEMonster, AlienVault, self-hosted ELK.

What is a Detection and Response Platform?

EDR / XDR / NDR

Endpoint Detection and Response (EDR) solutions provide in-depth visibility into the various devices deployed across the organization, allowing teams to detect and respond to any threats impacting them. EDR monitors endpoint events, analyzes processes and behaviors, and provides detailed forensic information to investigate and remediate security incidents.

Extended Detection and Response (XDR)s are designed to go beyond log analysis to provide a more holistic view of threats, correlating events across different security layers to uncover cyberattacks that might otherwise go unnoticed. The easiest way to understand XDR is to imagine that every endpoint on your network is also an Intrusion Detection System (IDS) sensor feeding all that raw data into an automated central SIEM-like correlation engine.

XDR: Pros and Cons

PROS	CONS
Proactive Threat Detection: Identifies and stops advanced threats that might bypass traditional antivirus.	Performance Impact: EDR agents can consume system resources and potentially impact endpoint performance.
Rapid Incident Response: Enables quick containment and remediation of endpoint threats.	Deployment Complexity: Deploying and managing EDR agents across a large number of endpoints can be complex.
Detailed Forensics: Provides rich forensic data to investigate and understand security incidents.	Cost: EDR solutions can be expensive, especially for organizations with many endpoints.
Improved Threat Detection: Can detect advanced threats and attacks that might bypass traditional security solutions.	Data Overload: Managing and analyzing the vast amount of data from diverse sources can be challenging.
Faster Incident Response: Automates and accelerates incident response, reducing dwell time and minimizing damage.	Integration Challenges: Integrating with various security tools and platforms can be complex.
Reduced Complexity: Simplifies security operations by consolidating multiple security tools into a single platform.	Vendor Lock-In: Due to the complexity and costs involved with setting up XDRs, it may be financially unfeasible to switch providers.

Network Detection and Response (NDR) solutions monitor network traffic for suspicious activity, using techniques like anomaly detection and threat intelligence to identify potential attacks. NDR provides visibility into network communications, identifies unusual patterns, and alerts security teams to potential threats.

NDR: Pros and Cons

PROS	CONS
Comprehensive Network Visibility: Provides a detailed view of network traffic and activity.	Data Volume: Analyzing large volumes of network traffic can be resource intensive.
Advanced Threat Detection: Detects sophisticated network attacks, including those that evade traditional security tools.	Deployment Complexity: Deploying and configuring NDR sensors can be complex.
Real-time Threat Monitoring: Enables continuous monitoring of network traffic for suspicious activity.	False Positives: NDR solutions can generate false positives, requiring investigation and tuning.

Sample Vendors to Research

- Sentinel One
- MS Defender XDR
- Palo Alto Cortex
- Darktrace
- Crowdstrike
- Trend Micro
- Microsoft Defender
- Extrahop
- Gigamon

What is SOAR?

Security Orchestration, Automation and Response (SOAR) platforms streamline security operations by automating incident response workflows. SOAR integrates with various security tools, allowing security teams to manage actions across different platforms, automate repetitive tasks, and manage incidents from a centralized dashboard. They are typically designed to democratize automation in a manner that does not require a highly experienced software development background. Using SOARs, organizations can focus more on strategic tasks.

SOAR: Pros and Cons

PROS	CONS
Increased Efficiency: Automates repetitive tasks, freeing up security analysts for more strategic work.	Initial Investment: Implementing SOAR can require a significant upfront investment in time and resources.
Faster Response Times: Accelerates incident response, minimizing the impact of security incidents.	Integration Complexity: Integrating SOAR with existing security tools can be complex and require ongoing maintenance.
Improved Incident Management: Provides a centralized platform for managing and tracking security incidents.	Skillset Requirements: Effectively utilizing SOAR requires specialized skills and expertise.

Sample Vendors to Research

- Torq
- Splunk Phantom
- Swimlane

What are TIPs?

Threat Intelligence Platforms (TIPs) aggregate and analyze threat intelligence from various sources, providing security teams with valuable context about threats, vulnerabilities, and attackers. TIPs collect threat data from threat intelligence feeds and analyze it for actionable insights.

TIPs: Pros and Cons

PROS	CONS
Enhanced Threat Visibility: Provides a comprehensive view of the threat landscape.	Data Quality: The accuracy and relevance of threat intelligence can vary depending on the source.
Proactive Security: Enables organizations to anticipate and mitigate threats before they materialize.	Integration Challenge: Integrating TIPs with other security tools can be complex.
Improved Decision-Making: Provides actionable intelligence to inform security decisions.	Cost: TIPs can be expensive, especially for organizations with limited budgets.

Sample Vendors to Research

- Flashpoint
- Crowdstrike
- RecordedFuture

What Are AI Models and Agents?

AI is rapidly transforming the security landscape, enabling organizations to automate tasks, analyze vast amounts of data, and proactively identify and respond to threats. AI models and agents play a crucial role in this transformation, offering various capabilities such as anomaly detection, threat prediction, vulnerability assessment, and automated incident response.

AI Model and Agents: Pros and Cons

PROS	CONS
Enhanced Threat Detection: Sometimes it is able to identify threats and anomalies that might evade traditional rule-based security tools.	Data Dependency: AI models require large amounts of high-quality data for effective training and performance.
Faster Response Times: Automates and accelerates incident response, minimizing the impact of attacks.	Bias and Fairness: AI models can be susceptible to bias, potentially leading to unfair or discriminatory outcomes.
Increased Efficiency: Automates security tasks, freeing up human analysts for more strategic work	Cost: Developing and deploying AI models can be expensive, requiring specialized expertise and infrastructure.

The Rise of Security Data Pipeline Management Platforms

Companies must ensure high data quality as they collect and manage security data. Challenges arise because data comes from potentially hundreds of unique sources with differing formats, structures, and of course, tremendous volumes of useless logs that are mixed in with relevant data. This requires cleansing, complex mapping, normalizing, or other processes to ensure accuracy and consistency, while making sure all stakeholders are happy with how the data is being handled, and how the organization is being protected.

Even if teams manage to figure all of that out, they are forced to painstakingly work with each single logging platform because of differing architectures, data mapping approaches, and analytics. Direct integration of data sources with SIEMs or XDRs can result in noisy data and a heavy onboarding or migration effort.

How do you solve the potential problems that might arise from integrating data directly with your SIEMs and other platforms? What if there was a “helper” or a translation layer between your data sources and data destinations that can take the heavy lifting of data operations off your internal team’s plate by decoupling the sources from destinations?

Security Data Pipeline Management (DPM) Platforms

A security data pipeline management platform helps decouple data sources from data destinations and adds the ability to operate on data before it reaches a destination. This removes individual onboarding dependency, and the prebuilt source and destination integrations make data easily routable.

Streamlined Quality Data

Filtering and data aggregation features can help reduce data. The DPM collects, reduces, enriches and routes data from various sources before sending it to analytics and storage platforms, improving the quality of data ingested at destinations.

Normalization and Enrichment:

A security focused DPM can allow for normalization of live-streaming data, integrations with threat feeds, applying enrichments on real-time streaming data vs post storage.

Dynamic and Context-Aware Routing:

DPMs allow for the dynamic routing of logs to multiple destinations, enabling the organization to split the stream based on predefined analytic use cases or specific security scenarios.

The DPM architecture decouples data sources from specific platforms, enabling the organization to seamlessly replace or augment SIEMs or other destination platforms without significant re-architecting. By allowing simultaneous data flow to multiple destinations, they can facilitate easy transitions to new SIEMs or cloud monitoring tools during migration periods, reducing integration costs and minimizing operational disruptions.

Sample Vendors to Research

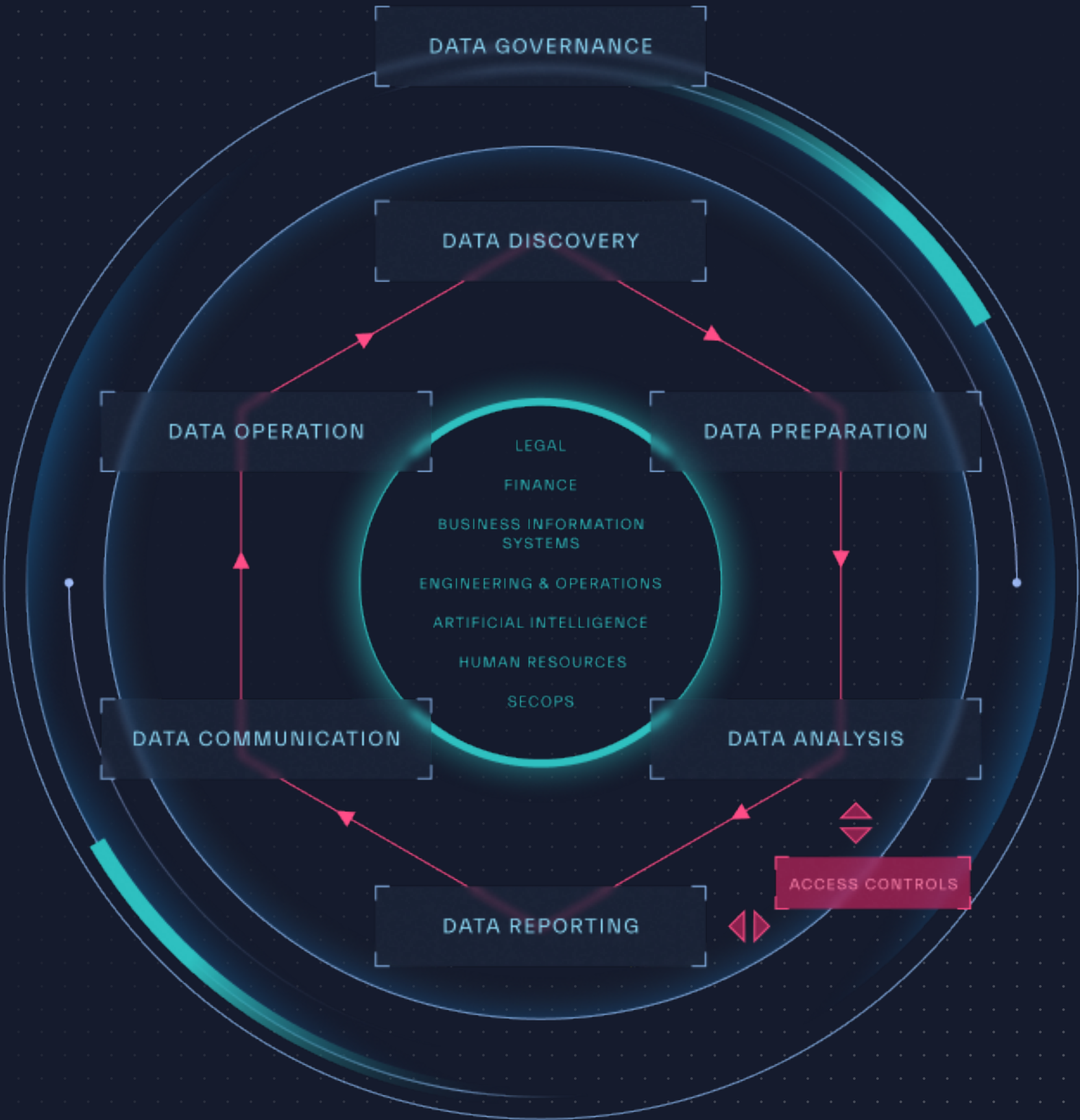
- Abstract Security
- Auguria
- BindPlane
- Fivetran
- DataBahn

Choosing the Right Tools for Your Needs

Selecting the appropriate data platforms is crucial for building a robust security posture. You need to consider your specific needs, security maturity, and budget when making your choice. Think of factors like scalability, integration capabilities, ease of use, interoperability, and vendor support.

Chapter 10

Time to Build Your Own Security Data Fortress



Time to Build Your Own Security Data Fortress

Now that you understand all of the essential components, you're equipped to build a data management framework using this Security Data Strategy Maturity Model. Using this approach, you can take control of your data to proactively identify and manage modern security threats.

Need a Refresher?

Here's what you should know about each pillar we've covered:



Recap: Data Discovery

Data discovery is the foundation of your security data strategy. It’s about identifying, cataloging, and analyzing all the data assets within your organization, creating a comprehensive map of your data landscape. This process is crucial for making informed decisions, improving operations, and ensuring compliance with regulations.

Key Takeaways You Should Remember

- 1. **Understanding Tactical vs. Strategic:** Data discovery initiatives can be driven by urgent, tactical needs like incident response or compliance audits. Conversely, they can be driven by long-term strategic goals such as cloud migration or creating new business units.
- 1. **The Five Phases of Data Discovery**
 - a. Source Identification
 - b. Initial Cataloging
 - c. Data Profiling and Sampling
 - d. Pattern and Keyword Discovery
 - e. System Ownership and Environment Mapping

Suggestions from Industry Leaders:

- 1. **Comprehensive Source Identification:** Ensure all known source types within your digital estate are treated consistently according to their data requirements.
- 1. **Automated Cataloging:** Mature data discovery programs leverage automated tools and metadata repositories, similar to API documentation tools, to facilitate data discovery and usage.

Recap: Data Collection

Data collection and ingestion form the first critical step in your security data pipeline. It's the process of gathering raw data from various sources and transporting it to a destination where it can be processed, analyzed, and stored. Efficient and reliable data ingestion is crucial for real-time threat detection, incident response, and proactive security measures.

Key Takeaways You Should Remember

- 1. **Identify and Tag Sources:** Start by identifying all relevant sources of security data, including logs from servers, endpoints, network devices, security tools, and cloud services. Tagging these sources with relevant metadata (e.g., data type, sensitivity, location) can improve organization and downstream processing.
- 1. **Optimize the Ingestion Pipeline:** Design your ingestion process to handle the increasing volume and velocity of security data without impacting system performance. Consider using a combination of push and pull mechanisms (e.g., Syslog, agents, API calls) to collect data efficiently.
- 1. **Filter and Aggregate:** Implement filtering and aggregation techniques to reduce data volume and improve efficiency. This helps focus on relevant security events and minimizes storage costs.

Suggestions from Industry Leaders:

- 1. **Standardize and Automate:** Implement standardized processes and automate data collection and ingestion tasks wherever possible. This improves efficiency and reduces manual effort.
- 1. **Prioritize Data Quality:** Incorporate data quality checks and filtering mechanisms to ensure that only relevant and trustworthy data enters your data pipeline.
- 1. **Plan for Scalability:** Design your ingestion pipeline to handle future data growth and evolving security needs.

Recap: Data Processing

Data processing is the refining stage of your security data pipeline. It involves transforming raw data into a usable format for analysis, reporting, and decision-making. This includes cleansing data to remove errors and inconsistencies, transforming it into a standardized format, and enriching it with additional context.

Key Takeaways You Should Remember

- 1. **Data Normalization:** Standardize data formats, including timestamps and time zones, to ensure consistency and facilitate analysis. Utilize common schemas like ECS, CIM, CEF, or LEEF for field mapping.
- 1. **Data Quality:** Implement checks and balances to continuously validate data quality and avoid structure drift. Monitor for data spikes, drops, and integrity issues.
- 1. **Evolving Architecture:** Adopt a flexible data processing architecture that allows for multiple consumers of processed data (data bifurcation) to support scalability and diverse use cases.

Suggestions from Industry Leaders:

- 1. **Enrichment Strategies:** Utilize data enrichment techniques to add context and value to your data but carefully scope enrichment usage to avoid unnecessary complexity and costs
- 1. **Aggregation and Summarization:** Employ aggregation and summarization techniques to reduce data volume without sacrificing essential information.
- 1. **Data Retention:** Establish appropriate data retention policies based on data type and business needs. Consider using a “flight recorder” approach for transient, high-volume data, retaining it only for a short period (hours or days) for immediate analysis and troubleshooting.

Recap: Data Analysis

Data analysis is the process of transforming raw data into actionable insights to understand, respond to, and predict security threats. It involves breaking down complex security challenges into manageable components and extracting meaningful intelligence from your data.

Key Takeaways You Should Remember

- 1. **Prioritize Business Needs:** Align your data analysis efforts with business objectives and prioritize security challenges based on criticality, risk, and impact.
- 1. **Iterative Process:** Data analysis is an iterative process involving hypothesis generation, data exploration, relationship modeling, pattern analysis, model design and evaluation, and continuous refinement.
- 1. **Actionable Insights:** Focus on extracting actionable insights from your data that can inform security decisions, improve threat detection, and optimize security operations.

Suggestions from Industry Leaders:

- 1. **Formalize Processes:** Establish standardized processes and methodologies for data analysis, ensuring consistency and repeatability.
- 1. **Invest in Tools and Technologies:** Utilize advanced analytics tools and technologies, including machine learning and data visualization, to enhance your analytical capabilities.
- 1. **Develop Expertise:** Invest in training and development to build data analysis expertise within your security team.

Recap: Data Reporting

Data reporting is the crucial bridge between data analysis and action. It involves communicating data-driven insights to decision-makers in a clear, concise, and visually compelling way, empowering them to take informed actions and drive organizational change.

Key Takeaways You Should Remember

- 1. **Audience-Centric Approach:** Tailor your reports to the specific needs and expectations of your target audience. Consider their roles, responsibilities, and the types of decisions they need to make.
- 1. **Visualize and Clarify:** Utilize data visualizations (charts, graphs, tables) to present complex data in an easily digestible format. Ensure reports are unambiguous and answer key questions relevant to the audience.
- 1. **Automate and Iterate:** Automate report generation and delivery to ensure timely and consistent reporting. Continuously refine and improve reports based on stakeholder feedback and evolving needs.

Suggestions from Industry Leaders:

- 1. **Gather Stakeholder Feedback:** Regularly solicit feedback from stakeholders to ensure reports meet their needs and drive desired outcomes. Use feedback to continuously improve report design and content.

Recap: Data Governance

Data governance and security are the guardians of your security data. They provide the framework, policies, and processes necessary to protect your valuable information assets, ensure compliance with regulations, and enable responsible data usage.

Key Takeaways You Should Remember

- 1. **Balance Oversight and Innovation:** Strike a balance between data governance and innovation. Implement policies and standards that mitigate risk without stifling growth and productivity.
- 1. **Centralize and Document:** Maintain a centralized access matrix and comprehensive data inventory to track data ownership, permissions, and applicable policies.
- 1. **Prioritize Data Security:** Implement robust security controls, including access controls, data masking, and deprovisioning processes, to protect sensitive data.

Suggestions from Industry Leaders:

- 1. **Formalize Policies and Standards:** Develop and document comprehensive data policies and standards that address data handling, access, retention, and security.
- 1. **Establish Clear Roles and Responsibilities:** Define clear roles and responsibilities for data governance stakeholders, including data owners, security teams, and compliance personnel.
- 1. **Implement Access Controls:** Implement robust access controls to ensure that only authorized individuals can access sensitive data.

Seeing it all Work Together

If your current security data management strategy is lacking in a specific pillar, you should start there and then augment it. For example, let's say you just noticed an unusual spike in network activity. What do you do? Well, your process should look something like this:

- Data discovery has already identified the relevant network logs.
- Your collection and ingestion pipelines bring that data into your processing hub.
- The data is normalized and enriched.
- The enriched data is stored in a hot tier for rapid access.
- The data is analyzed further, suggesting a potential intrusion.
- A report is generated which alerts your security team(s) and decision-makers.
- SecOps and DFIR teams engage further to interdict the threat.
- Governance policies dictate your response.

Taking each component into consideration, where do you stand? How mature is your organization? Compared to your current framework, were any steps lacking? If so, identify the gaps and ask yourself: **how are these gaps impacting our ability to detect and respond to threats?**

Perhaps your data discovery process is incomplete, which is creating blind spots that you can't afford to ignore. Maybe your storage strategy isn't optimized, leading to slow query times and increased costs. Are your reporting mechanisms failing to deliver timely insights to the right people?

By grading your organization using this maturity model, weaknesses and improvement opportunities can be identified to enhance and augment your security data strategy.

Appendices

All documents are available here:

[Data Strategy Supporting Documents](#)



Appendices

A. Asset Inventory Workbook

Download here: <https://abstract.security/book/asset-inventory>

B. Access Matrix Workbook

Download here: <https://abstract.security/book/access-matrix>

C. Policy and Standards Template

Download here: <https://abstract.security/book/policy-standards>

D. Log Onboarding Checklist

Download here: <https://abstract.security/book/log-onboarding>

E. White House Executive Order Regarding Event Logging

Download here: <https://abstract.security/book/exec-order>

Appendices

F. Matt Carothers SCTE TechExpo24 – Stupid Log Tricks

Download here: <https://abstract.security/book/matt-carothers>

G. Equifax “Safety in Numbers” Patent

Download here: <https://abstract.security/book/safety-in-numbers>

H. SigmaHQ Translation Backends

Download here: <https://abstract.security/book/sigmahq-translations>

I. SigmaHQ Rule Repository

Download here: <https://abstract.security/book/sigmahq-rules>

J. Backup Policy Document

Download here: <https://abstract.security/book/backup-policy>

Appendices

K. **Analysis Use Cases**

Download here: <https://abstract.security/book/analysis-use-cases>

L. **Julien Pileggi - SIEM Rules You Should Implement Today**

Download here: <https://abstract.security/book/siem-rules>

M. **Adventures in Data Labeling**

Download here: <https://abstract.security/book/adventures>

What Readers Are Saying



Nick Mullen

CEO, Entoo Security



"Good data security doesn't happen by accident, and it doesn't happen at all without an executable strategy. This guide does a great job of explaining the what's, why's, and how's of designing a robust data security program, with a blueprint that is easy to understand and follow."

Andrew Borene

Executive Director, Flashpoint and Former Senior Officer, the Office of the Director of National Intelligence



"Applied Security Data Strategy" is an essential read for anyone seeking to understand the evolving landscape of data management. This book stands out because it's authored by a powerhouse team of industry leaders who provide a clear, practical roadmap for developing a security-driven data strategy. With step-by-step guidance, it equips professionals with the tools needed to navigate today's complex threat environment while ensuring both resilience and operational effectiveness."

Melinda Marks

Practice Director, Cybersecurity, for Enterprise Strategy Group



“Our modern world revolves around using software applications in both our personal lives and for work. While it makes it easy and convenient to do everything from shopping, to paying bills, to working, the amount of data exchanged – across everything from laptops to mobile devices, servers, and networks – continues to proliferate. Also, as software development continues to increase productivity and scale with technological advances including AI, organizations need to be diligent about needed security processes and controls to mitigate risk to protect their data as well as customer data. This book provides a useful guide from industry experts on data management and what is needed for effective security.”

Rick McElroy

CEO, Nexasure



"In an era where data is the new currency and cyber threats evolve by the second, securing information is no longer optional, it's survival. This book unveils the strategies and mindset needed to protect what matters most."

Francis Odum

Founder, Software Analyst Cybersecurity Research



“Over the next decade, the role of data in the SOC will fundamentally shift from being an underutilized asset to the foundation of security operations. Today’s SOCs struggle with data visibility gaps, misaligned telemetry, and the high cost of storing and analyzing security logs. To truly optimize SIEMs and SOC strategy, security engineering teams must take control of their data—understanding what they collect, why they collect it, and how it drives better detection and response. This isn’t just about cost reduction; it’s about reducing risk, improving signal-to-noise ratios, and ultimately enhancing the experience for SOC analysts. The future of the SOC will belong to those who treat data as a strategic advantage rather than an afterthought. This book provides most of the foundational structure and framework of any security team to support them along that journey.”

What Readers Are Saying

"This book provides a useful guide from industry experts on data management and what is needed for effective security."

Melinda Marks, Practice Director,
Cybersecurity, for Enterprise Strategy Group

"The future of the SOC will belong to those who treat data as a strategic advantage rather than an afterthought. This book provides most of the foundational structure and framework of any security team to support them along that journey."

Francis Odum, Founder,
Software Analyst Cybersecurity Research

"This guide does a great job of explaining the what's, why's, and how's of designing a robust data security program, with a blueprint that is easy to understand and follow."

Nick Mullen, CEO,
Entoo Security

"This book unveils the strategies and mindset needed to protect what matters most."

Rick McElroy, CEO,
Nexasure

"With step-by-step guidance, it equips professionals with the tools needed to navigate today's complex threat environment while ensuring both resilience and operational effectiveness."

Andrew Borene, Executive Director,
Flashpoint and Former Senior Officer, the
Office of the Director of National Intelligence

