# SODEF23: The Cost-Effective Security Information & Event Management (SIEM) Framework: How to Build Threat Detection Systems While Saving Resources

Benjamin Rader
IUPUI
Purdue School of Engineering & Technology

**Abstract**—This paper delves into the critical domain of cybersecurity, with a specific focus on the challenges and methodologies surrounding Security Information and Event Management (SIEM) systems. It acknowledges the integral role of cybersecurity in modern business infrastructure, emphasizing its evolution from an optional feature to a fundamental necessity for safeguarding operations. The discussion pivots around the intricacies of managing complex IT systems and the imperative of balancing simplicity with effectiveness in IT architectures. The core contribution of the paper is the introduction of the Security Operations Data Engineering Framework (SODEF), a comprehensive guide designed to streamline security data management. SODEF is dissected into various critical components, including IT Systems, Data, Data Pipelines, and others, with a special emphasis on the pivotal role of data pipelines in enhancing the efficiency and cost-effectiveness of security operations.

## 1 Introduction & Literature Overview

Security is a crucial aspect of today's problem landscapes, and business operations, and has manifested in the cyber world as cybersecurity. When not accounting for the ill intentions of external actors, businesses eventually fail to operate for their intended purpose. In other words, cybersecurity is now a requirement to be a business. When cybersecurity is absent, businesses lose compliance, shareholders, marketability, and control of their systems. However, cybersecurity is not necessarily as tactical as it is painted in the media. Sometimes cybersecurity is merely designing the business to lower risks on the cyber front. Oftentimes, properly established cybersecurity in an organization will involve the detection of threat actors on IT systems. Without detection or visibility, organizations are blind and cannot prioritize security initiatives or redirect resources to incidents or problems.

The Internet is an abstract collection of systems. On a small scale, creating networks that can solve business problems is straightforward. Use data available to the business in combination with those systems to compute, interface with the business's employees, and ultimately allow for informed decision-making. The issues start to appear as these information technology networks get larger. Due to the abstract design of IT systems, larger architectures quickly gain complexity and incur costs. There is a delicate tug-of-war between complexity and effectiveness in IT operations. One design principle that relates to this is the idea of minimizing the amount of allowed outputs for a system while maximizing the amount of allowed inputs. Often, the hardest problems to solve with IT are the ones that have overengineered architectures or a vast number of interacting systems and interfaces, protocols, and data to go along with them. Simplified architectures work the best. In this paper, I attempt to show how security operations architectures can be designed and simplified to improve threat detection.

### 1.1 Security Information & Event Management (SIEM)

Detecting threats or events with the potential to cause harm if they happen [1], is difficult with IT systems because they do not all speak the same "language" and they are sometimes hard to find or keep track of. Asset management or tracking what systems exist is a problem that often takes a dedicated team to handle in an organization. On the security and threat detection side, finding threats on these systems is difficult as well, and it is the focus of this paper. A metaphor could be used to describe the nature of this problem. When doing threat detection, the business's network could be viewed as a large hospital with patients being the IT systems or nodes on that net. The process of log management and threat detection would be akin to what doctors are doing. The log management part would be the doctors conversing with the patient, noting down the information, and then transporting it to other systems for processing. Doctors observe patients and examine them for any issues or symptoms. They may send patient data off to some resource to process the data and then make conclusions about the patient. This is akin to threat detection. The problem of log management and threat detection or analysis becomes more apparent by adding one more circumstance to the hospital. Imagine that, on average, only one in fifty people speak the same language. In a hospital where it is hard

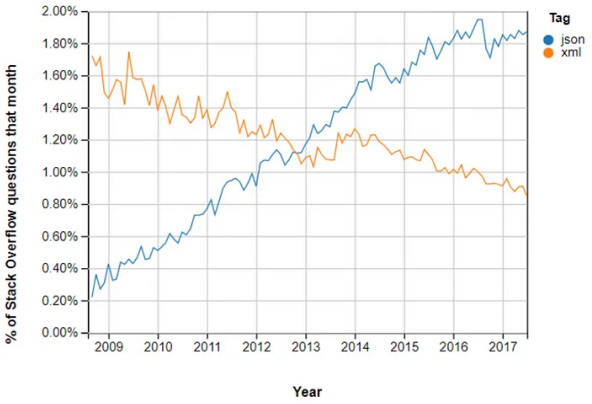


Fig. 1. Popularity of XML vs JSON based on Number of Questions on StackOverflow

to find people who speak the same language, it becomes difficult to extract information from patients or run the data through analysis systems, because they may not be used to certain languages. This illustrates the problems being dealt with in IT with threat detection and analysis across a myriad of IT systems. Security information and event management (SIEM) platforms and log analytics platforms are the manifestation of a solution to this problem.

Below is a definition of SIEM from Microsoft:

> "a solution that helps organizations detect, analyze, and respond to security threats before they harm business operations. (...) SIEM technology collects event log data from a range of sources, identifies activity that deviates from the norm with real-time analysis, and takes appropriate action. In short, SIEM gives organizations visibility into activity within their network so they can respond swiftly to potential cyberattacks and meet compliance requirements. In the past decade, SIEM technology has evolved to make threat detection and incident response smarter and faster with artificial intelligence."

Logs from IT systems are also not sufficient to effectively respond to threats. Typically, security operations teams need OSINT (open source intelligence) data, curated lists of certain indicators such as IP addresses, and other points of data to utilize alongside their IT systems and the logs they produce [2]. SIEMs are changing and some claim that they are growing increasingly irrelevant or ill-equipped to handle the modern scale of data and lack of talent to engineer solutions that can utilize the data [3]. To simplify these problems, security teams must understand the landscape enough to create systems for threat detection or integrate existing ones into their infrastructure.

### 1.2 Data & Protocols

The primary complicating factor of threat detection and event/log management architectures is the presence of various data formats and protocols when it comes to log-producing IT systems. The hospital metaphor for IT systems and disparate languages of IT systems would be a monumental issue if it were not for established standards for communications. The Request for Comments (RFC) series is one of the most well-known publications for Internet standards. Among the standard-setting bodies for the Internet, the Internet Engineering Task Force (IETF) is the most prominent [4]. Jon Postel, one of the most decorated writers of RFCs, has been called the "Editor of the Internet [5]." Over the years, data formats have been part of the RFC series. Today, two data formats stick out more than any other and can cover most use cases of log and event data. Those data formats are the tabular or flat format "CSV" (comma-separated values) and the nested JSON (Javascript object notation) file format. In terms of tabular formats (think Excel files), the CSV file has been around since the 1970s. The CSV format was made official by RFC-4180 [6]. Tabular formats are incredibly useful for transporting relational and structured data. Where this structure is not possible, JSON is the next option. However, some logs or events do not follow either of these formats. Some may follow a delimited format which can be nested or flat such as a log using a mix of colons and brackets. Extensible Markup Language (XML) used to be the most popular option for nested data. However, over the years JSON has been adopted as the simpler and practical option for nested data (shown in Figure 1.) JSON was first introduced in 2001 but grew quickly in popularity as big tech companies like Google and Facebook started using it. XML also had additional security risks in implementation, so REST APIs and JSON became standard for development [7]. On a side note, formats like Parquet are good for optimizing with certain use cases such as where one column in the tabular format is being used. Parquet performs better than CSV in those cases, so there are some caveats to CSV and use with flat data. Based on these facts, a good approach to threat detection would be to utilize data in these two formats as much as possible with some outliers where optimization is worth it. This makes it easier to transport and more likely to interface well with data engineering infrastructure and threat detection setups.

### 1.3 Purpose & Problem Statement

This paper is an attempt to define the problem areas of threat detection systems from the perspective of data engineering (managing data and its compatibility and usability) in a way that allows for the cost-effective design of a SIEM. Currently, there are no frameworks like this that define a taxonomy for components that make up a SIEM. Even if there were, such a taxonomy is not being used in the industry. As a result, many SIEM companies make solutions that are considerably different from other platforms with the same marketing buzzwords. Some SIEM tools can only ingest security data from particular sources and are limited in analysis capabilities. Conversely, some enterprise SIEM platforms include data normalization logic, and data pipelining, or utilize a custom query language like in the Splunk platform. SIEM tools are a sort of black box that has not been completely defined by their components. In this paper, I will attempt to segment those components into an architecture that could be the ultimate definition of a security information and event management system. By modularizing the idea of SIEMs, other tools can be combined or integrated to fill certain gaps. The goal of this

framework is to ease the process of designing SIEM systems which saves on costs such as administration efforts, license costs, and hardware usage among other line items.

## 2 Security Team Research

### 2.1 SIEM Importance & Usage

During research, it was important to look at the stats on SIEM usage in organizations. One pattern that became obvious was that SIEMs are expensive to implement properly without the right talent, but they have great benefits and data to back up their efficacy. For instance, the "Cost of a Data Breach Report" from IBM in 2023, showed that organizations utilizing SIEMs, on average, saved over $200,000 in the event of a breach. Additionally, teams with AI or machine-learning-driven insights, including those from a SIEM, tended to save $225,000 more during a breach. Additionally, teams who detected a breach with internal tools tended to save, on average, 1 million per breach [8]. Another survey specifically for SIEMs surveyed 348 cybersecurity professionals. It found that 56% of respondents utilized a SIEM, hybrid deployment adoption was rising, 84% reported reduced occurrence of breaches with SIEM, 81% said it improved threat detection, and those who used a SIEM reported higher levels of confidence in their security posture ( %15 more.) Those who used a SIEM defined several benefits of their use: efficient sec-ops, faster detection and response, and better visibility into threats. Most reported that they could detect threats within minutes or at least in hours. Some of the systems teams were seen integrating with them include: IDS/IPS (intrusion detection and prevention), next-gen firewalls, application logs, EDR, and ransomware detection or anti-malware controls. 71% of teams reported monitoring and correlation of activity across multiple systems as their top use case. More than half of the teams saw threat intelligence integrations as consequential for these systems [3]. Ironically, some would not define SIEMs as always necessitating threat intelligence data and it appears to be quite important to many teams. This shows a divide between expectations and what has traditionally been offered as selling points. Despite the mass benefits that SIEMs come with, they also come with a swathe of challenges.

### 2.2 Cybersecurity Operation Budgets & SIEM Costs

Security operations teams can have it tough when it comes to control implementation in an organization especially if the team is immature. Without a large enough budget, most security teams are forced to rely on a narrow set of controls. USTelecom surveyed more than 300 small and medium-sized enterprises (SMEs) engaged in critical infrastructure work. Growing security operations teams had the most budget by almost triple (Figure 2) [9]. This stat included mostly small businesses, but it goes to show that growing security operations programs need fuel. A big problem with these budgets is that they are typically only around 10% of the IT budget despite cybersecurity touching the majority of the revenue. That budget share is mostly made up of staff and compensation with only about 30% of it for on-prem and cloud software in which SIEM tooling would fall [10].

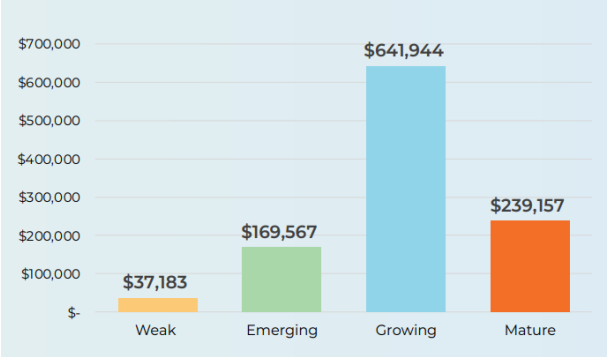


Fig. 2. Cybersecurity cultural segments with average budget. 75% of respondents had less than 100 employees. Includes 300 SMEs.

One survey of nearly 1,600 different CISOs (chief information security officers) showed that the top 3 priorities with these budgets are as follows: 1) innovate, 2) consolidate controls, solutions, and tooling, and 3) outsource the security controls. Last are handling insider threats and awareness training [11]. SIEMs are one great way to innovate if implemented correctly. However, it is hard to implement such controls when most security operations teams have so little room to wiggle. If the cybersecurity budget were a pie, then SIEMs are equivalent to having to figure out how to fit twenty apples into the pie while still having room for other fruits. To fit them, one would likely need specialized equipment (specialized employees [3], [12]) to break down the apples to their purest form. Such is the dilemma of fitting a SIEM into the budget. The team needs to fit many other tools into their budget, to the point where the SIEM tooling only takes up 14%

Most security operations teams in 2020 have outsourced at least one security service. The most outsourced security function, by far, was the SIEM. Granted, SIEMs sometimes need to operate around the clock [12]. When teams move to new SIEMs, they are not doing so because they primarily want a managed service, better detection algorithms, and alerting functionality. Companies move to new SIEM vendors because of cost (%16.8.) The other reasons seem to revolve around usability and innovation [13]. It is a chore to reap the tool's benefits before the costs catch up with the team while executives ask for some proof of ROI (return on investment.) All in all, if organizations want value out of a SIEM, there are many hurdles to overcome: finding skilled staff (%41), filtering out false positives (%37), and a lack of budget (%34) [3]. If most teams do not have room for SIEMs in their budget, but most teams have a SIEM, then the individual implementations must be examined for a better understanding of how this contradiction exists. In other words, it would seem that most teams can afford a SIEM and see their price as the biggest pain point.

## 3 SIEM & Data Landscape Research

### 3.1 SIEM Tools

With a simple Google query, one will find at least 20-30 SIEM tools. With so many of them, it is apparent that they are targeted at different markets and use cases. The market is largely led by Splunk, Microsoft Sentinel, and

IBM Qradar. Many customers utilize analytics platforms that are ”cloud-native.” Several key features can be found in SIEM solutions: real-time security monitoring, UEBA (user and entity behavior analytics), data visibility, incident management, threat intelligence, mapping to compliance standards, and security orchestration automation and response (SOAR.) Tools like Splunk are seen differently through the lens of data analysts and engineers. Many view Splunk as a monitoring tool of sorts or only for log analytics. However, as any data engineer could recognize, SIEMs are not the only tools that can analyze logs and integrate with other functionalities like a full-fledged tool such as Azure Sentinel.

### 3.2 Data Engineering & Data Pipelines

Engineers are professionals who invent, design, analyze, build, and test machines, systems, and structures. In the past, this referred to physical systems, but the definition has evolved to be applied to all sorts of topics, ontologies, and applications. With IT, some engineers work with these IT systems. Often, architects are the ones who find the problems and map them to the business. The engineers are the personnel who figure out how to implement some of the technology on a deeper level. Engineers work with engines and that means a lot of moving parts. Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale. SIEMs work primarily with data, but as was mentioned, they usually include lots of supplementary content to make analysis easier and give security operations teams an easier time implementing threat detection with the data they have. Analytics tools have a hard time working with this data as it is, so data engineering systems must be included in these SIEM tools, in analytics platforms, or implemented separately to alter the data into a viable state. Data pipelines are the primary way in which data is transformed in this way.

> A data pipeline is a method in which raw data is ingested from various data sources and then ported to a data store, like a data lake or data warehouse, for analysis. Before data flows into a data repository, it usually undergoes some data processing. This is inclusive of data transformations, such as filtering, masking, and aggregations, which ensure appropriate data integration and standardization [14].

In the solution landscape of data engineering, data pipelines are incredibly valuable. Data pipelines are extensions of some methods that are central to data operations. Another name for these is ”ETL” or extract-transform-load tools. The primary objective of data pipelines is to transform data and give it more usability and compatibility with other systems. However, there are arguably other objectives that are just as critical such as data reduction and aggregation or consolidation of data during the pipeline. All of these processes in the data pipeline rely on how hardware is related to costs and convenience. By far, the most important aspect of data pipelines is that they typically operate only in memory. Other ways to describe this are stream-based data processing or stateless data processing. Essentially, this means that each collection of data such as an event or log is processed separately and the system does not need to keep track of ”state.” This results in only needing to operate in memory unless the events themselves require more memory than RAM provides. Nevertheless, this allows for data engineers to continuously ”stream” data through these pipelines and catch the data coming out the other side. The benefits of this approach can be related to the same idea of how routers and stateless firewalls operate.

These data transformation pipelines can reduce data to save money on computation in analysis systems or save time for indexing systems that organize the transformed data into data warehouses. The point of these pipelines is to simplify and optimize the data to make other data engineering processes cost-effective and valuable.

### 3.3 SIEM Marketing & New Solutions

One interesting pattern of these SIEM systems is that the tools do not all do the same thing. This is interesting in the fact that a lot can be gleaned about security operations data management from the diverse set of solutions. Some tools like Wazuh or Graylog seem to be marketed at small businesses that want an all-in-one solution that can be implemented in a day. Some SIEM systems only work with security data for servers. Many enterprise SIEMs have more flexibility in the analysis that can take place, but they fall short when it comes to efficient data engineering and the standardization and reduction control that come from the use of data pipelines. The most comprehensive event management systems such as Splunk can work with all sorts of data. Still, they are usually cost-prohibitive and require the consolidation of multiple data-using teams into one tool [2]. This can be a good thing, but it should not be necessary for cost-effectiveness. Unfortunately, organizations will need to spend nearly as much time as an SIEM research project takes to figure out which tool they can use and they will likely need to make compromises.

When marketing solutions, the definitions, buzzwords, or names they use will often change. The changes are often of little significance, but rather a strategy to communicate novelty or get more attention. Marketing is not the focus of this paper, but many solutions have the same features, functionality, and power that SIEMs do while going by a different moniker.

### 3.4 Security Data Pipelines & Security Data Lakes

The hardware and backend challenges that come with mapping hardware to storage architectures to analytics platforms are massive. This is the same for SIEMs too. By going with an SIEM, teams have to sacrifice efficient or cheap backends and the license costs suffer. If a security team tries to use an optimized backend through an analytics platform, then they make concessions with the lack of integrations for the processing of security-related data. Big data platforms and companies began marketing big data analytics platforms for security in this way back in the 2010s. These were the birth and death of the first security data lakes [15].

Data lakes are a ”subset” of data warehouses. They can store unstructured, semi-structured, and structured data. They are a place where anyone can throw any type of data

quickly and simply. Amazon Web Services S3 service is an example of this. Another huge benefit of these storage services is that they are cost-effective by nature. Users can define the type of storage they want based on how often they will do reads vs writes. As a result, one can easily get storage that only costs a fraction of a cent per month per GB. This is incredibly cheap especially when compared to the cost per GB for ingestion pricing models with SIEM tools. One of the most popular SIEMs called "Sentinel" from Azure charges $296 per day for 100 GB ingestion per day. That equates to $2.96 per GB per day. Considering storage and ingestion are separate, the numbers still tell a lot about the pricing of data storage versus that same full markup with SIEM tooling. Due to this fact, many SIEM companies have gone all in at integrating these new cloud services with security-focused analytics platforms. These architectures are more modular and require lots of thoughtful components, but they are more attainable in the 2020s due to the "SaaS-ification" of all these complicated data backends. This is the birth of the security data lake and security data pipeline, and by 2023, they are finally coming to fruition [15].

Security teams typically did not have time to implement security projects like Apache Spot or Metron, but now SIEM providers are outsourcing parts of their backend to cloud service providers and allowing customers to "bring their own data lake." This also saves organizations money by reducing egress fees associated with cloud service providers. Security data lake tools bring the same content for analytics but allow organizations to use their storage. This means that organizations will need to implement ETL or data transformation pipelines to put the data into these data lakes though. Enter the security data pipeline.

There are several purposes for having a security data pipeline [16]:

1) Log Visibility - all systems that can send logs can be managed from one place
2) Analytics Platform or SIEM Portability - easy to move data to a new analytics platform or SIEM by changing the pipeline's destination. It could be a nightmare if one needs to reconfigure an army of agents
3) Data quality
    - Keeps detection engineers (analysts who write detections in threat detection systems) and content engineers from having to remove noise using ill-equipped SIEM tools (not generally built for data transformation)
    - Logs that are not normalized have higher time-to-value with additional storage/compute resources demanded at the SIEM layer
4) Expensive Data Collaboration & Inefficient Routing
    - The SIEM deployment will need to be scaled for more users if other teams need access to the same data or forwarding systems will need to be implemented separately for access to the data.
    - Custom data backends mean for sometimes difficult access processes to the raw data
    - Forwarding raw logs may need to be done from the original log source whereas a consolidated pipeline would make redirection and forwarding less of a headache

Some examples of modern security data pipelines include Tenzir, Cribl Stream, and Tarsal. Arguably, Cribl Stream has the most value because it seems flexibly applied to more than security use cases and could consolidate the data quality and log visibility operations for multiple teams. Most of these tools, if not all of them, are cloud-native as well which works great for saving on cloud service fees and other caveats. Security operations teams should consider refactoring their threat detection systems into a more modern and cost-effective architecture. Yet, security data pipelines will not always be the answer. Security teams need a consistent framework to fit their tools into, and that is what I have developed through my survey of the data landscape of threat detection systems.

## 4 SODEF: The Security Operations Data Engineering Framework

### 4.1 Relating SIEM Components to Data Engineering

I have designed a framework for viewing the problems of security information and event management and more broadly as a lens for data engineering with regards to security operations. This framework is outlined in the diagram in Figure 3 (page 6.)

The traditional view of a SIEM is quite limited. This was partly a motivation for this project. As noted in section 3.1, SIEM tools have a variety of implementations in today's market. It is tricky to define what a SIEM system entails. Therefore, I have curated most of the SIEM tools on the market and reviewed their components along with modern approaches to event management with solutions like security data pipelines and security data lakes. These are merely marketing terms and do not define anything concretely. Therefore, I ventured to make a flexible framework that can help to explore the design of such systems in a way that is cost-effective in the new cloud and data-driven economy where security operates.

I will call this framework the **Security Operations Data Engineering Framework** or "**SODEF**" for short.

Components of the Cost-Effective SIEM Framework (Fig. 3 – pg 6):

1) **IT Systems:** These are the systems that produce events or logs. They do not necessarily have to be IT systems either. Sometimes these could be manual entries from SecOps personnel. A caveat to IT systems is that they will use any number of protocols or processes to transfer data to other systems in the security data operations architecture. Many times, agent-based collection systems will be employed to collect logs from these IT hosts, computers, and/or software and send them off to the next node. There can be much nuance to how this data is transmitted depending on the network topology, level of network security required, and TCP/IP protocols

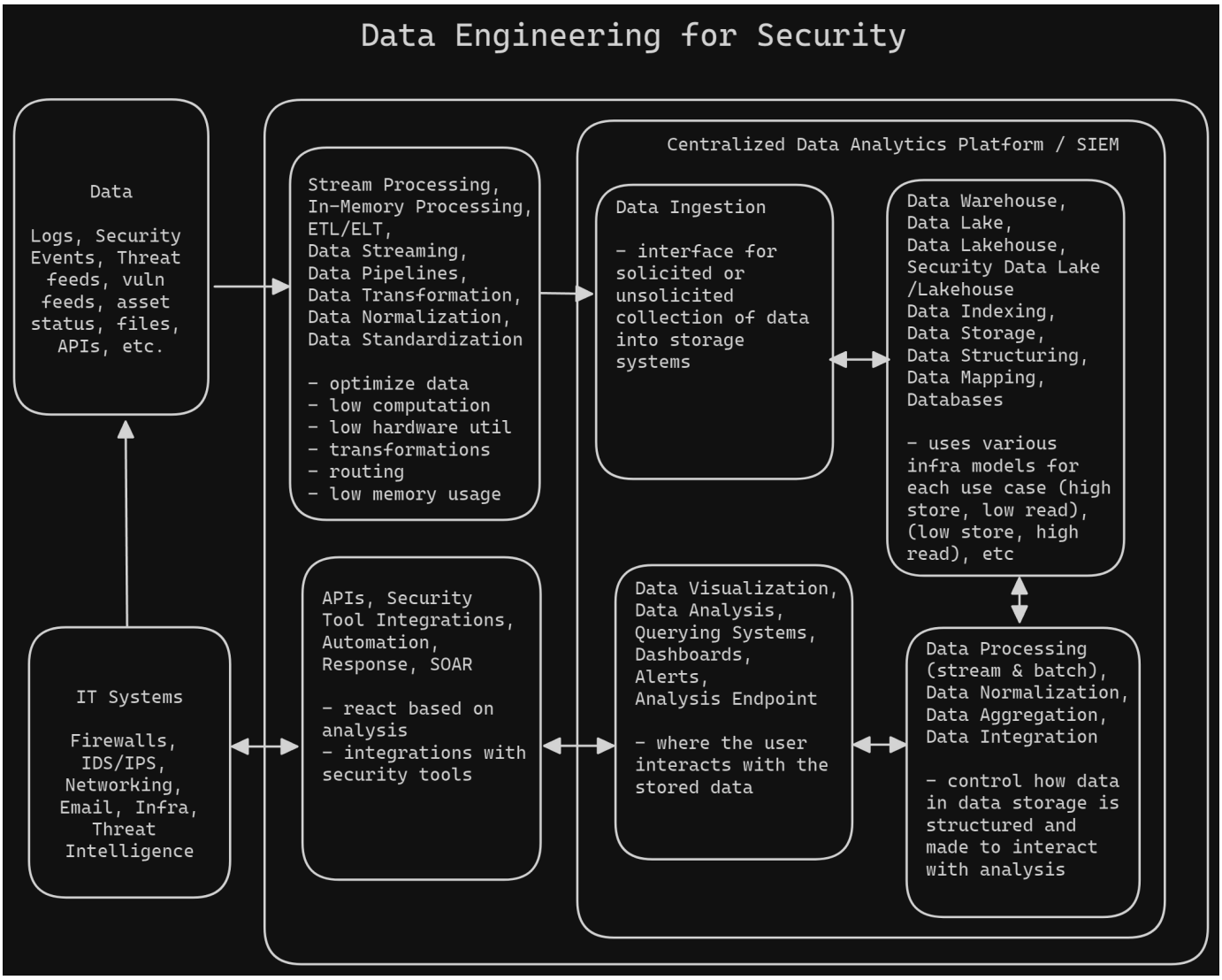


Fig. 3. **DIAGRAM - The Security Operation Data Engineering Framework (SODEF):** breaks down the problem areas of security data management and the building of threat detection systems into 8 interrelated categories. In traditional terms, this is a cost-effective framework for building or implementing SIEM systems.

involved in the exchanges. Not to mention hardware must also be closely accounted for at this level. Examples of this include firewalls, intrusion detection and prevention systems, networking devices, infrastructure (in general), and threat intelligence producers such as a honey pot.

2) **Data:** Data is dimensional, dynamic, and heterogeneous. It can come in many forms, shapes, and sizes. This is a crucial piece for security teams to think over. If the data is relatively structured or there is high confidence that the data will come in a certain format, then other areas of the architecture such as data pipelines for data quality may not need to be as comprehensive. If the data is quite unstructured, then that knowledge can inform other components during system design as well.
3) **Data Pipelines:** Data pipelines should broadly refer to a place or proxy where most event data is funneled or directed. Having one place to send all logs makes routing later in other steps. However, this step should also be used for consolidation of data quality operations, data normalization or standardization, observability operations, and data transformation. Another common aspect of this node should be that of low computation and in-memory processing per log or event. This means that the processing should not be stateless where the processing of one log relies on the results of having processed logs before it. However, even if this is not the case, it should usually happen all within memory. Data Pipelines are not the only marketed solution that can accomplish this piece of the SIEM puzzle. Other ways to define this step in the process are data transformation, data normalization, data reduction, and ETL processes, proxies, or middleware.
4) **Centralized Analytics Platform or SIEM:** This is merely a way to define what most SIEMs or data analytics platforms will handle. Most will not handle data pipelining or have SOAR (security orchestration, automation, and response) baked into them.

However, SIEM tools will normally have a way to ingest data, a storage system, a way to process or analyze the stored data, and then visualization or a system for getting content or value from the processing of the data.

5) **Data Ingestion:** This part is not essential to the framework. However, it is common with SaaS or cloud-based SIEM tooling. Analytics platforms require functionality in place to have secure unsolicited and solicited transfer of events into a storage system such as a data lake or lakehouse. This could be complex depending on where the warehouse, SIEM, data lake, or other logical storage silo is hosted. Perimeter security and general IAM (identity access management) would ideally be implemented in something like a VPC (virtual private cloud.) Therefore, how data is transferred across the board is important, and it is best to simplify this piece as much as possible since costs will not be affected as much as a node like the data lake or warehouse.
6) **Data Lake/Lakehouse/Warehouse:** The purpose of this node is to store or index the events in a way that optimizes infrastructure costs and data usability during analysis. SecDataOps (security data operations) personnel should capitalize on synergizing the analytics users' use cases with the cost optimization of the data stores. Data pipelines are also intimately involved with data storage by choosing which types of storage to put events into (cheap-to-read vs. cheap-to-write.) An example of this would be the infrequently accessed tier of S3 buckets from AWS. They are cheap to store data in but will incur more costs if users need to read from them. This could happen if the security team "archives" logs that the security team or incident response (IR) may need during an investigation. If the logs are stored in an organized way, then costs can be optimized still by having the retrieval be specific enough to only read a small slice of the warehoused or indexed data.
7) **Data Processing:** This is the node that is used to interface the data store and analytics nodes. When data is analyzed, it may need to be done in batches or as a continuous stream. This is the node that accomplishes that. Although this abstract component will usually not be implemented by the security team, it can affect the compatibility between analysis and the data in the warehouse. For instance, a custom system may need to be implemented to package the data up or aggregate certain data points before being analyzed by the analysis node. It is important to account for the integration of the data store with the analytics piece of security operations.
8) **Data Analysis:** This node allows the security operations team to finally get value from their data by using anything from data querying engines, security dashboards, visualization libraries, and even machine learning or artificial intelligence models. At this node of the framework, security data operations engineers should understand what questions the security operations team has and how to answer them with a threat detection system. This is also where detection engineering would take place. Security engineers can correlate processed events with other events to make certain conclusions such as the existence of anomalous user behavior on a network.
9) **Response & Automation:** This includes the use of APIs, external tool integrations, and SOAR tools (security orchestration, automation, and response.) This layer responds to content and results from the analysis node. An example of this could be to automatically block malicious IPs found during the analysis phase.

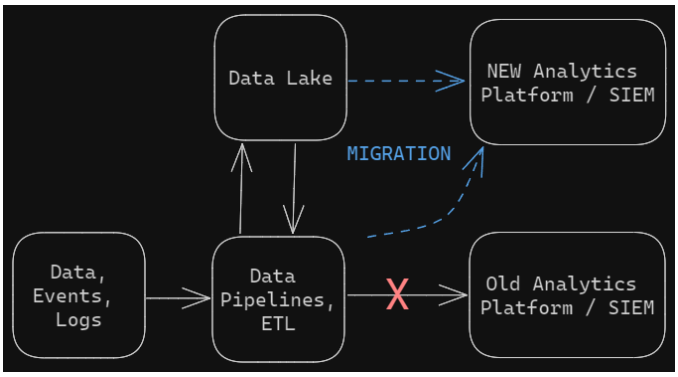


Fig. 4. Diagram illustrating the portability that comes with the implementation of data pipelines in a security operations data engineering stack.

### 4.2 Maximizing Cost Efficiency in the SODEF Framework: the Pivotal Role of Data Pipelines or ETL

The most cost-effective way to utilize this framework is to start with a system for building data pipelines. Data pipelines are useful because data quality and cost-optimization can be drastically improved. Yet, with many threat detection system undertakings, they have the major advantage of being the primary place for all events to be sent. This becomes the collection point for all logs. This has the outcome of enabling future migrations out of limited enterprise SIEMs to other modular and other accurately-intentioned architectures. The administrative overhead of going into hundreds of IT systems and configuring the rerouting of logs can take months of monotonous work, IAM involvement, and bureaucratic bottlenecks. In past SIEM implementations without an approach that meets this framework, SIEMs had to be used and could not perform normalization, parsing, and event transformation tasks that are essential to data quality and usability of data.

There are plenty of tools that are marketed to accomplish this step for security teams while integrating well with this modern framework including Cribl Stream, Tenzir, Tarsal, and even more telemetry-focused solutions like Mezmo. The paradigm shift that is taking place with these solutions has been motivated by security team frustrations with analyzing events for threats with the largest frustration being the cost of these systems. Data pipelining and ETL tools enhance an organization's precision of log utilization, leading to a reduction in infrastructure expenses through the deliberate optimization of storage and computation resources. Most

SIEMs are exorbitantly expensive because they lack optimization when it comes to data warehousing. By using data pipelining tools, security teams have control over how their data is organized, and instead of warehousing all of the data, they can warehouse what they want with tools like chaos search, use specific ”partitioning” schemes for their data to have organization to them, and they can use an analyze on read approach, where the data does not take up resources unless it needs to be analyzed. In most modern enterprise SIEMs, loads of data are indexes and warehoused when they will never get used and this results in huge unnecessary expenses.

Data pipelines also have the added benefit of simplifying SecOps and log management. By using pipelines, all logs can be focused on the ETL or pipelining system no matter where they come from. Teams do not have to question where to send logs and it means better consistent processes. Since an ETL tool is being used, it is also likely that there already exists a myriad of pre-made interfaces for IT systems that have complicated interfaces for the transmission of the logs. Such interfaces can usually be found under ”integrations” of a similar name for the solution. Interfaces and actual content such as transformation, reduction, and normalization logic are typically packaged with the ETL tool based on use cases they are marketed towards. For security data pipelines, the integrations will likely apply to systems like firewalls, cloud service provider applications, and middleware common to log ingestion tasks such as OpenTelemetry.

Another strength of ETL is that it simplifies the work of migrating SIEMs (Fig.4). Once any normalization and transformation logic is moved from the SIEM into the ETL tool, the process for migrating to a new analytics platform is as simple as attaching the analytics to your storage (data lakes) or pointing the ETL tool at the SIEM.

The last added advantage of using ETL in security operations is that it can be implemented to allow for better data quality and visibility while troubleshooting upstream conflicts. Upstream conflicts refer to changes that take place in a process early on which affect any subsequent logic, operations, and systems. In this case, the upstream errors that occur in an ETL tool will relate to IT system changes such as with networking or data format changes which could happen if new log-producing are added or an update occurs that changes the shape or nature of the data. Data quality and error handling or checking can be implemented as a part of the in-memory processing that takes place in the data pipeline. This allows for almost instantaneous detection of upstream conflicts, data format changes, or the chance of data loss downstream. There are data quality and ”observability” tools that can accomplish this task with better efficiency. However, the consolidation of this process into ETL can save teams chunks of their budgets.

In conclusion, there are four prominent benefits to having ETL as a part of one’s security operations data engineering architecture:

1) **Cost optimization** related to infrastructure computation and storage
2) **Simplified log management and routing** tasks for IT systems and data owners (partly due to network topologies and employee collaboration nuances)
3) **Simpler migration** out of data analytics or SIEM platforms
4) **Consolidated data quality processes**, error handling or checking for upstream conflicts, better observability, easier troubleshooting, and enhanced log visibility

### 4.3 Caveats to Reduction with Data Pipelines

Data pipelines are incredibly valuable for security teams to implement because they align data with the goals of the analytical systems. However, there are some caveats to building systems that have data pipelines. The largest caveat is that proper data transformation pipelines and the building of such architectures require skills and knowledge related to infrastructure management, networking, and data engineering skills such as an understanding of programming, query languages, virtualization technologies, and familiarity with cloud service provider models. Finding an employee who can be a ”security champion” for DevOps (developer operations), data engineering, and data analytics can be difficult. It may be rare for security engineers, let alone analysts, to have these sorts of skills while also having the time to implement or design these systems. On the other hand, it may be hard to communicate or emphasize security with a DevOps or data engineer when it comes to these systems. There will be tradeoffs and caveats to the implementation of these systems concerning the availability of specialized or capable human resources.

### 4.4 Framework Scope, Use Cases, Threat Model

The SODEF framework can benefit teams of any size. However, the security budget of the organization and its threat model will greatly affect how the architecture should be realized. For instance, a small company that only uses an application or two for security or user data may benefit more from a monolithic SaaS offering that has fewer features and control over the data. The risks may be minimal and threat detection with ML or AI would likely not be worth the cost for a small team. On the contrary, a large team with hundreds of applications and over ten security tools would likely save hundreds of thousands by implementing a security data ETL process or data pipeline attached to optimized storage. It is also likely that a modular implementation would benefit other IT operations teams as well [17]. When it comes to synthesizing threat detection architectures, organizations should first focus on understanding their threats. Then, the data engineering system should be based on the threat model and risk appetite of the organization.

## 5 Data Pipeline Implementation

I’ve implemented a real-world data transformation pipeline by using the ”Cribl Stream” ETL tool. Below are an example of the transformed logs. Including the original logs is unnecessary in illustrating the value. The transformation pipeline was tested over 1000 events which came out to an %18.44 reduction in average event size. An example or result of one of these events being reduced is shown in Listing 1 (below.) This shows an example where keys have been ”flattened” so that keys are not nested. In other

words, there are not multiple levels of keys and values. This transformation into flattened JSON also involved the dropping or filtering of many key:value pairs from the original data, so that only the necessary data is included. Flattened JSON is useful because it's very easy to convert to a tabular format such as the CSV format. Additionally, it is easy for analytics platform users to instantly understand the data.

Event size is not the only reduction that can be done either. Typically, the two reductions which will take place is a reduction in event size and a reduction in event throughput or the events that are kept. Events that are not kept are simply dropped. This is easy and efficient to do with in-memory processing. With these two reductions alone organizations can save enough money on their budget to buy a medium-sized home.

### 5.1 Results & Savings Analysis from Pipeline Reductions

A simple Cribl Stream pipeline implementation which involved dropping fields and flattening the JSON data, results in an average of %18.44 reduction in event size. To show the value of a seemingly small reduction such as this I will show the cost savings across three popular SIEMs' pricing models, and then I will add in an additional %60 reduction based on organizational data that I obtained. In a distributed Kubernetes environment in an organization with hundreds of developers and numerous teams, well over %60 (probably realistically closer to %80) were logs at the "informational" level. These logs are generally not supposed to be needed in production environments. Therefore, I will test a hypothetical 2,000 Gigabytes daily ingest and 90,000 Gigabytes stored logs across the following SIEM pricing models: Elastic Cloud, Microsoft Sentinel, and Splunk.

Starting with Splunk, they use a stock of computational usage called the SVC or "Splunk virtual compute." It measures how much computation has been used and changes depending on how often the logs being stored are expected to be used. The storage is somewhat dynamic and that means if you have 10 SVCs and only use logs for compliance, then beginning to do "exploration" type activity with the data can put a security team above their entitled license [18]. A realistic or conservative example of the cost of an SVC could be $15 per day per SVC unit. One interesting pattern with SVCs is that part or even most of the cost will be linearly related to the amount of data ingested. Assuming %50 of total SVC usage is affected, the savings can be calculated based on the ingest component. An organization that ingests 2000 GB daily needs about 114 SVCs. This ultimately equates to $624,150 after a year and $312,075 related to ingest. The result is %16.3 of the original cost with both the reduction in event size and dropping info logs.

With both event size & throughput reductions: $312,075 * SIZE_REDUCTION(1-.1844) * THROUGHPUT_REDUCTION(1-0.8) = $50,905.674 (after reductions)
$261,169.32 in savings.

With only event size reduction:
312,075 * SIZE_REDUCTION(1-.1844) = $254,528.37
$57,546.63 in savings

Other estimates are hard to find. In fact, only a few platforms provide straightforward estimates. The Elastic Cloud estimates pricing calculator results in anywhere from $500,000 to $4,000,000 depending on the array of "hot", "warm", and "cold" storage is used [19]. Generally, a mix will be used with each being expected to be used more often than the next. Cold storage is usually only read or used during an incident or catastrophic issue. It's difficult to get objective estimates presently since all of the platforms use different methods for mapping hardware usage to costs. Nonetheless, these costs show the savings that can be had when using an ETL tool. Below are the optimal savings if logs were reduced %20 by size and %50 by throughput with a 4 million dollar license.

With both event size & throughput reductions: $4,000,000 * SIZE_REDUCTION(1-0.2) * THROUGHPUT_REDUCTION(1-0.5) = $1,600,000 (new cost)
$2,400,000 in savings with a %60 reduction in data being used by hardware in the analytics platform.

Lastly, I have the ingest-based pricing from Azure Sentinel. This model is simple but tends to have high prices since the SIEM providers must assume a fixed amount of hardware usage to likely profit. In other words, most of these pricing models bank on users not utilizing all of their data just as Costco relies upon consumers who fail to make good use of their memberships. Ingest models are priced in a way that will only be cost-effective if all of the logs are analyzed and utilized. At 2,000 GB of ingest per day, the price is $4,800 per day which comes out to $1,752,000 over a year [20].

With both event size & throughput reductions: $1,752,000 * ALL_REDUCTION(0.4) = $700,800 (new cost)
$1,051,200 in savings with a %60 reduction in logs taking up computation.

## 6 Conclusion

The financial implications of adopting data pipelines in cybersecurity operations, as delineated in the Security Operations Data Engineering Framework (SODEF), are profound and far-reaching. By analyzing various SIEMs, the framework demonstrates potential savings from hundreds of thousands to over a million dollars. Such financial benefits are not merely incremental; they represent significant budgetary relief for security operations teams. For instance, a million-dollar saving can drastically augment a team's capacity for innovation, enable the acquisition of advanced security tools, or fund critical staff expansions. These savings are pivotal in a domain where budget constraints often limit the effectiveness and scope of cybersecurity measures.

This enhanced efficiency and cost-effectiveness, highlighted through the implementation of data pipelines, is a testament to SODEF's practical value. The application of tools like Cribl Stream or other ETL tools exemplifies how event size reduction and strategic data optimization

```
{
  "host": "ip-##-###-###-##.ec#.internal",
  "message": "{\"@timestamp\":\"####-##-##T##:##:##.###-##:##\",\"sequence\":########,
  \"systemClassName\":\"org.faker.logmanager.system\",\"systemName\":\"stdout\",\"level\":
  \"INFO\",\"message\":\"INFO  com.atc.openauction.sql.ClientInfoDataSourceAdapter
  {fcid=#c#a#e##-####-####-adea-###ea#dbbfdd, span_id=###dd###df#abdc#, trace_flags=##,
  trace_id=####fc##e#a#a##ebc#fe#e##bff##c#}:  Proxy or Dynamic generated class found:
  org.faker.threads.EnhancedQueueExecutor$ThreadBody\",\"threadName\":\"default task-###\",
  \"threadId\":######,\"mdc\":{},\"ndc\":\"\",\"hostName\":\"ab-standard-ps-#fbf#b###d-xtcvs\"
  ,\"processName\":\"faker-modules.jar\",\"processId\":###,\"@version\":\"#\",\"log-handler\":
  \"CONSabE\"}",
  "docker.container_id": "####b#bb#ccba#a##ab###a#c##a##c####f###f#bcb##########d#ef##d#b#",
  "kubernetes.container_name": "ab-standard-ps",
  "kubernetes.namespace_name": "example#-prod",
  "kubernetes.pod_name": "ab-standard-ps-#fbf#b###d-xtcvs",
  "kubernetes.container_image": "artifacts.secops.io/sec/example#/ab-standard-ps:#######",
  "kubernetes.pod_ip": "##.###.##.###",
  "kubernetes.host": "ip-##-###-###-###.ec#.internal",
  "kubernetes.labels.app_kubernetes_io/name": "ab-standard-ps",
  "kubernetes.namespace_id": "##c##fa#-ecfa-#d#d-#b##-#######dd#d#",
  "level": "info",
  "hostname": "ip-##-###-###-###.ec#.internal",
  "log_type": "application"
}
```


Listing 1: Optimized log after Cribl Stream data pipeline. Notice the JSON has no nested keys or all the fields are "top-level." This is real-world data and has been anonymized and modified.

translate into measurable financial gains. These are not theoretical advantages but real, quantifiable improvements in operational expenditure. Additionally, data quality can be improved so that events can be more readily available to SIEM users, content producers, and analysts.

Despite these benefits, adopting such advanced approaches in cybersecurity comes with its challenges. The necessity for specialized skills in data engineering and a profound understanding of cybersecurity is paramount. This necessitates aligning the framework's implementation with an organization's specific threat landscape and risk tolerance, underscoring the importance of a tailored cybersecurity approach.

In sum, SODEF advocates for a significant shift in cybersecurity operations. Transitioning from traditional, monolithic SIEM systems to a more nuanced, flexible, and financially viable framework is critical. Central to this framework is data pipelines, which not only confront current challenges but also lay a foundation for adaptable and robust cybersecurity operations amidst evolving threats and technological progress. The tangible financial benefits gleaned from the implementation of data pipelines provide a compelling argument for organizations to reevaluate and realign their cybersecurity strategies for enhanced efficiency and effectiveness.

## REFERENCES


[1] I. Muscat, "Cyber threats, vulnerabilities, and risks — acunetix," Acunetix, 08 2019. [Online]. Available: https://www.acunetix.com/blog/articles/cyber-threats-vulnerabilities-risks/

[2] D. Schoenbaum, "The average siem deployment costs 18m annually…clearly, its time for a change!" Medium, 11 2022. [Online]. Available: https://schoenbaum.medium.com/the-average-siem-deployment-costs-18m-annually-cf576f6c740d

[3] C. Security, "The state of security: Siem in 2022 — tripwire," www.tripwire.com, 08 2022. [Online]. Available: https://www.tripwire.com/state-of-security/state-of-security-siem

[4] "Request for comments," Wikipedia, 11 2023. [Online]. Available: https://en.m.wikipedia.org/wiki/Request_for_Comments#:~:text=The%20RFC%20system%20was%20invented

[5] P. B. B. OBE, "Ode to rfcs and jon postel," ASecuritySite: When Bob Met Alice, 09 2021. [Online]. Available: https://medium.com/asecuritysite-when-bob-met-alice/ode-to-rfcs-and-jon-postel-22af34fd3b85

[6] M. Drapeau, "The state of csv and json," Medium, 10 2018. [Online]. Available: https://medium.com/@martindrapeau/the-state-of-csv-and-json-d97d1486333

[7] K. Parmar (KPBird), "Rfc 8259 — the javascript object notation (json) data interchange format," Medium, 06 2019. [Online]. Available: https://kpbird.medium.com/rfc-8259-the-javascript-object-notation-json-data-interchange-format-607f70fd1

[8] IBM, "Cost of a data breach 2023," IBM, 2023. [Online]. Available: https://www.ibm.com/reports/data-breach

[9] USTelecom, "Ustelecom 2023 cybersecurity culture report," USTelecom, 02 2023. [Online]. Available: https://www.ustelecom.org/research/2023-cybersecurity-culture-report/

[10] L. Columbus, "Benchmarking your cybersecurity budget in 2023," VentureBeat, 02 2023. [Online]. Available: https://venturebeat.com/security/benchmarking-your-cybersecurity-budget-in-2023/

[11] Proofpoint, "2023 voice of the ciso — proofpoint us," Proofpoint, 05 2022. [Online]. Available: https://www.proofpoint.com/us/resources/white-papers/voice-of-the-ciso-report

[12] J. Vijayan, "35 stats that matter to your security operations team," TechBeacon, 2021. [Online]. Available: https://techbeacon.com/security/35-stats-matter-your-security-operations-team

[13] P. Security, "State of siem report 2022," Panther Labs, 2022. [Online]. Available: https://panther.com/resources/reports/state-of-siem-2022/

[14] IBM, "What is a data pipeline — ibm," www.ibm.com, 2021. [Online]. Available: https://www.ibm.com/topics/data-pipeline

[15] O. Singer, "Why your security data lake project will succeed!" Medium, 10 2022. [Online]. Available: https://osinger.medium.com/why-your-security-data-lake-project-will-succeed-3f6484d17b3

[16] A. Teixeira, "Why you need data engineering pipelines before an enterprise siem," Medium, 10 2023. [Online]. Available: https://detect.fyi/why-you-need-data-engineering-pipelines-before-an-enterprise-siem-0be553584aa9
[17] Splunk, "Get the report: State of security 2022," Splunk, 2023. [Online]. Available: https://www.splunk.com/en_us/form/state-of-security.html
[18] ——, "Pricing calculator," Splunk. [Online]. Available: https://www.splunk.com/en_us/products/pricing/pricing-calculator.html
[19] ElasticCo, "...," cloud.elastic.co. [Online]. Available: https://cloud.elastic.co/pricing
[20] M. Azure, "Azure sentinel pricing — microsoft azure," azure.microsoft.com. [Online]. Available: https://azure.microsoft.com/en-us/pricing/details/microsoft-sentinel/
[21] Microsoft, "What is siem? — microsoft security," www.microsoft.com, 2023. [Online]. Available: https://www.microsoft.com/en-us/security/business/security-101/what-is-siem
[22] M. Ghaida, "Json: A brief history, and a look into the future," Medium, 01 2021. [Online]. Available: https://markghaida.medium.com/how-to-set-up-an-api-fetch-request-in-rails-3798ad10f079
[23] A. Singh, "The story of json: Simplifying data like never before," Medium, 08 2023. [Online]. Available: https://medium.com/@mrarunsingh8/the-story-of-json-simplifying-data-like-never-before-a768bd70c6fc
[24] SecurityMagazine, "Report shows cybersecurity budgets increased 6
[25] Cribl, "Bracing for impact: Why a robust observability pipeline is critical for security professionals in 2023," Cribl, 02 2023. [Online]. Available: https://cribl.io/blog/why-a-robust-observability-pipeline-is-critical-for-security-professionals/
[26] I. News, "Tenzir's security data pipeline platform optimizes siem, cloud, and data costs," Help Net Security, 08 2023. [Online]. Available: https://www.helpnetsecurity.com/2023/08/09/tenzir-security-data-pipeline-platform/
[27] B. Rader, "Jsonaut," GitHub, 12 2023. [Online]. Available: https://github.com/cybersader/jsonaut