LAKE IT OR LEAVE IT

The case for owning your security data

DATABAHN

CASE STUDY

In 2015, a publicly-traded insurance company purchased a SIEM solution, which we'll refer to as SIEM15. Three years later, due to the inadequate threat detection capabilities of SIEM15, the company invested in a second SIEM solution, SIEM18. By 2021, facing financial constraints and finding SIEM18's UEBA capabilities insufficient, they switched to another SIEM solution, SIEM21, which provided both UEBA and SIEM functionalities. Despite transitioning through 3 different SIEM vendors, the company is still obligated to pay full renewal fees for all 3 solutions - SIEM15, SIEM18, and SIEM21 - due to regulatory requirements mandating the storage of 3 years' worth of data. Consequently, they cannot retire SIEM15 until 2025, and will face similar challenges with SIEM18 until 2028.

Having had negative experiences with previous solutions, the company requested that SIEM21 write data into their own data lake. However, they now face two significant challenges: the data in their data lake is not in an open-source format, creating difficulties for threat-hunting teams. Each SIEM they've used has its own querying language, complicating the learning process for the teams.

The SIEM licenses, either based on daily data ingestion or compute usage, have limited the CISO's budget and have prevented the inclusion of all telemetry data in the SIEM. As a result, the key raw traffic data, like EDR and firewall logs, is stored in the Enterprise data lake. This fragmentation has slowed threat hunting, forensics, investigations, and analysis.

Moreover, the SIEM21's data lake had tables written into it in a sub-optimal format, leading to substantial infrastructure cost and effort due to the need of large querying instances. Changing their SIEM solution would alter the data format and could prevent data from populating their data lake, as they heavily rely on the SIEM's ingestion process.

> **It feels like we have non-functioning leased cars parked in our garage just because they open the garage door flawlessly.**
>
> **Head of SOC,** *publicly-traded insurance company*

The volume-based pricing models and compliance requirements are combining to create onerous constraints of SOCs and CISOs, actively preventing them from focusing their resources and bandwidth towards improving their security, but more focused on security data logs and volume.

SIEMs & SECURITY DATA

A Brief History

Understanding why SIEM vendors continue to offer certain capabilities requires a look at their history. In the early days of SIEM, with platforms such as Nitro Security, Q-Radar, and ArcSight, security technologies were primarily on-premises. These systems featured a simple connector framework, a logger (using RDBMS databases), and an ESM that performed basic threat detection rules, such as brute-force attack identification. As cloud-based SIEMs emerged, vendors retained some legacy components like loggers, which were tightly coupled with the threat detection components.

Why was this architecture favored?

- **They wanted to avoid routing sensitive data to the enterprise data warehouse**
- **They needed SIEMs with data lake capabilities to store data for compliance & forensic purposes**
- **They required robust threat-hunting capabilities**

Post 2019, platforms such as Snowflake, Databricks, AWS Security Data Lakes, Microsoft Azure, Google Cloud, and Elastic have revolutionized data storage with scalable and flexible infrastructure. Most enterprises have either adopted a comprehensive cloud-based data lake strategy, or are in the process of transitioning to one. IT teams now possess the skills and expertise to effectively manage these platforms.

Given these advancements, does it still make sense for cybersecurity teams to rely on a data lake strategy that is tightly coupled with the SIEM? There are 2 factors influencing customers to own their own data lake, and why this is the right time for them to make that change - **market changes** and the pressing need to **improve efficiency while reducing redundancy**.

Market Changes

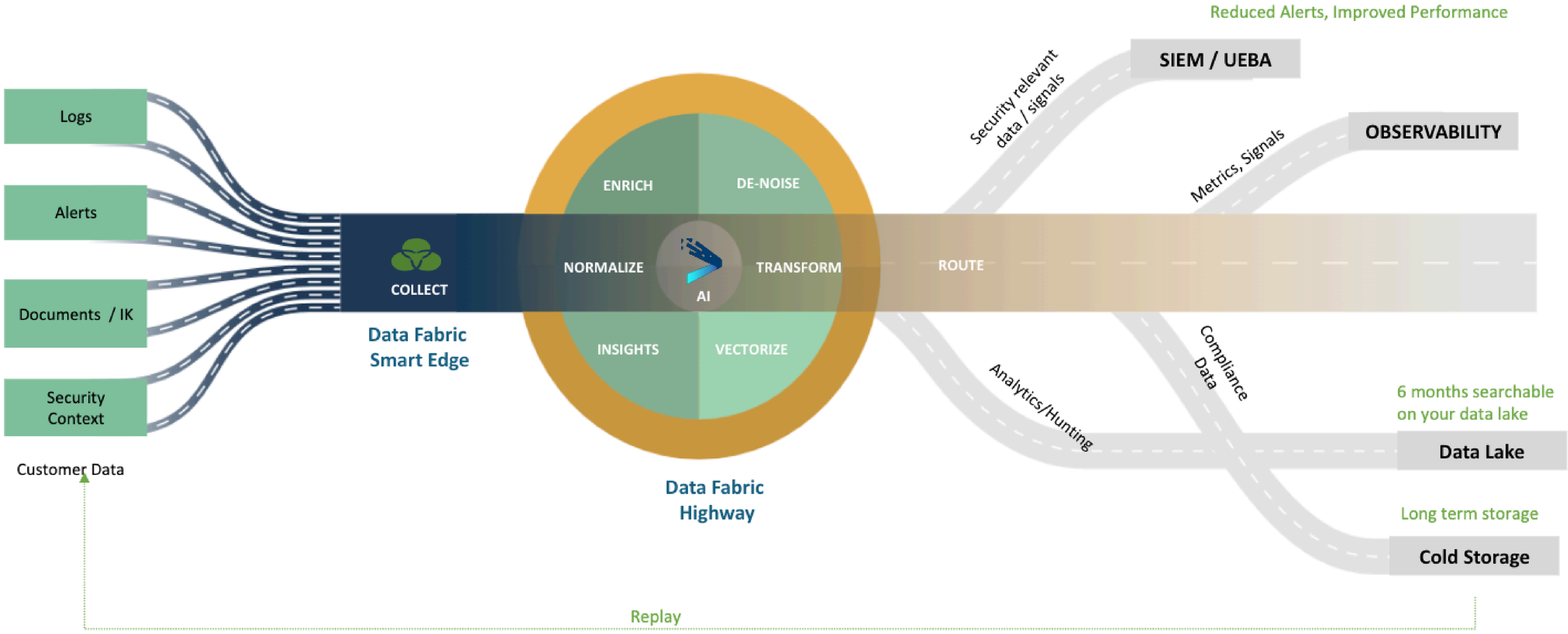
- **Recent M&A and consolidations:** 3 out of the 5 vendors in the 2024 Gartner Magic Quadrant for SIEMs have been affected, significantly altering the market landscape. This creates uncertainty for existing customers, where their teams might have to navigate obsolescence or degradation.
- **Costs and pricing:** SIEM customers could also face increased pricing pressure with industry consolidation and a reduction in the number of large enterprise vendors; especially with the high switching costs associated with migration when data collection and SIEMs are coupled.
- **Customer needs:** The need for better threat detection and response from SIEM vendors is leading to very few SIEM vendors offloading ingestion and storage responsibilities due to the high costs, low margins, and the lack of importance to SOC teams.
- **The search for Search:** SIEM vendors have not delivered optimal search experiences due to a lack of focus on that functionality; especially when contrasted against Data Lakehouse / Warehouse products which are purpose-built for such use cases and have made significant improvements.

This has led to a significant cohort of leading enterprise customers to architect their data lakes with solutions such as Databricks, Snowflake, and AWS Data Security Lake to significantly enchance search performance, providing faster and more reliable data retrieval.

Improving efficiency and reducing redundancy

- **Dealing with multiple data storage endpoints:** Many organizations store data in multiple locations. For example, infrastructure teams need Unix and Windows logs for monitoring, licensing, and timely patching, while cybersecurity teams need the same logs for threat detection.
- **Relevant and Irrelevant data:** SIEMs are usually tightly coupled with data collection in enterprise security stacks, and end up ingesting both security-relevant and security-irrelevant data, leading to redundancy.

- **Unified pipeline:** A unified pipeline product or a centralized data ingestion platform can separate and write security-relevant and non-security data to different tables in the enterprise data lake and can significantly reduce costs, network bandwidth consumption, and increase efficiency.
- **Planning and building for AI**: Some organizations are ready or building their readiness for **RAG** (Retrieval-Augmented Generation) of the data on their own data lake to obtain deeper, AI-driven insights on their data.



CONCLUSION

To address the challenges, the publicly-traded insurance company is now writing data into their own data lake, powered by AWS and Snowflake. They are using DataBahn's Data Fabric solution to populate the data in the Open Cybersecurity Schema Framework (OCSF). The search performance on their AWS / Snowflake setup is markedly better.

For enterprises using legacy SIEMs, own your own data lake. It gives you the flexibility to adapt to industry changes, improve search capabilities, and maintain control over your data. The modular approach helps you manage and upgrade your security infrastructure effectively and efficiently, ensuring robust protection in an ever-changing threat landscape. It unlocks customized security measures, data sovereignty and compliance with regulatory requirements, and control over where and how your data is processed and stored.

ABOUT DATABAHN

DataBahn.ai's Data Fabric empowers organizations to optimize data management, reduce costs, and enhance security and IT operations. By integrating AI readiness, addressing IT observability challenges, and offering flexible solutions, the platform delivers significant operational efficiencies and strategic benefits, setting a new benchmark for cybersecurity data management in the digital age.

Learn more at **databahn.ai**

DATABAHN