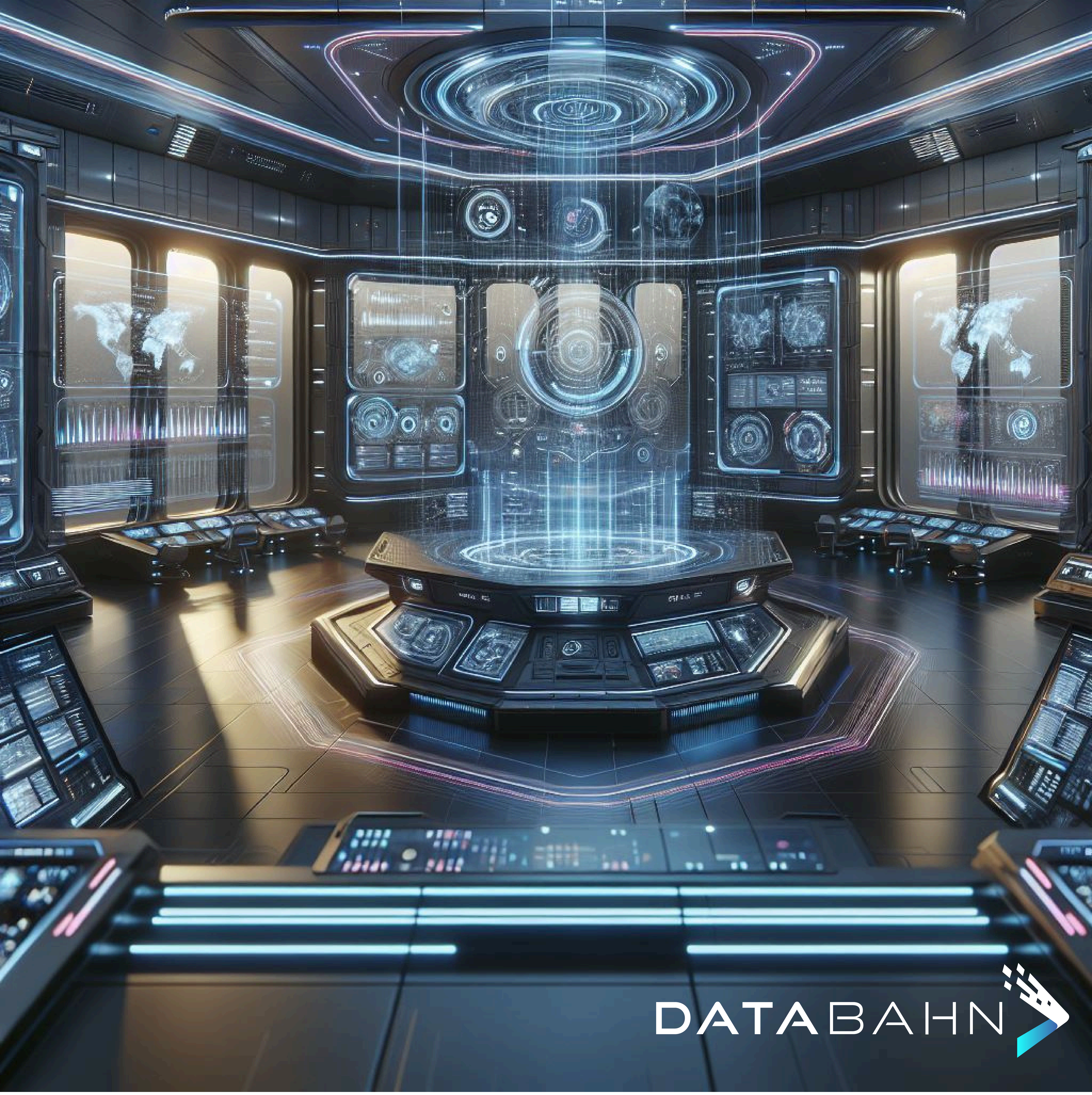


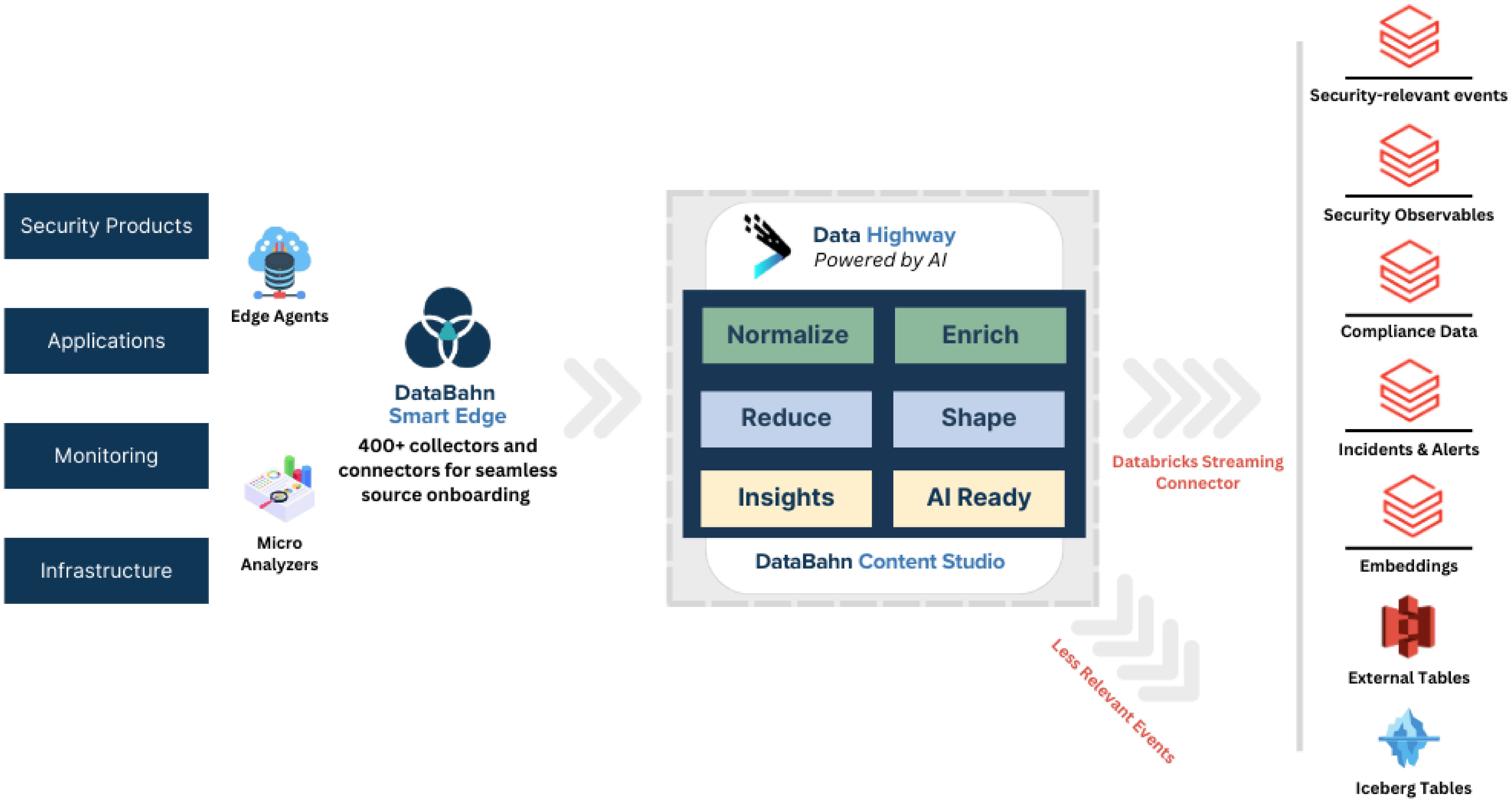
SOLUTION BRIEF

Building the Next-Generation Security Data Lakehouse in Databricks using DataBahn's Security Data Fabric

DataBahn for Security Data Orchestration in Databricks

Databricks is being chosen by many enterprises to be their data lake. Databrick's Data Lakehouse allows for complex data science use cases, analytics, and ML operations with Its **ManagedMLflow** offering. Databricks is also noted for its ability to scale and process a large amount of data and has support for multiple languages with extensive libraries - which makes it favored by large production enterprises in complex industries, especially when there are use cases for multiple data types. Databricks also offers a separate storage layer which is independent of its processing layer, and can act as an ETL tool to add structure to the unstructured data.

Databricks can scale up to meet the high throughput demands of any high-volume system, and has extensive support for continuous writes and concurrency. However, SOCs face challenges in managing custom data pipelines to centralize log ingestion, ensuring query performance, and unpredictable costs. This has led to many security teams choosing not to realize the value of migrating from monolithic SIEMs to data lakehouse powered SIEMs like Databricks.



The Solution

DataBahn helps Databricks users by streamlining data collection and ingestion and removing the burden of building customized integrations and customized pipelines, deploying staging locations, or managing your data orchestration to send relevant data for analytics and processing, as well sending less-relevant events to external tables and iceberg tables.

DataBahn's purpose-built Smart Edge along with the Data Highway platform can take data from a range of sources (both on-premise and cloud), parse and structure them into any format or data model of your choosing, enrich data with any meaningful context (internal and external), orchestrate the data to extract meaningful insights and deliver data and insights into Databricks for optimal querying, high performant analytics and search, thereby reducing your overall operating costs with Databricks.

Through DataBahn's orchestration capabilities, SOCs and Security Teams can:

- **Simplify data collection and ingestion into Databricks**
    - Using DataBahn's plug-and-play integrations and connectors with a wide array of products and devices
    - Using DataBahn's native streaming integration for a hassle-free, real-time data ingestion into your Databricks deployment
    - By effectively normalizing and structuring complex data using DataBahn's orchestration pipelines before the data is loaded into Databricks tables
- **Convert logs into insights**
    - By using volume reduction functions like aggregation and suppression to convert noisy logs like network traffic / flow into manageable insights that can be loaded in Databricks reducing both the volume and the overall time for queries to execute
- **Increase overall data governance and data quality**
    - Identify and isolate sensitive data set in transit thereby limiting exposure
- **Perform split-second threat hunting**
    - Use DataBahn's indicator index to extract insights such as Security Observables (IP addresses, Domains, URLs, Hashes), Entity Relationships (Processes, Network execution, Registry modifications), and Intel Context
    - Use additionally derived context such as first observed / last observed time / frequency of observation to speed up data exploration and hunting
- **Use best-of-breed services and technologies**
    - Leverage DataBahn's simplified data orchestration capabilities, Databricks customers can use additional tools to implement a cyber mesh architecture without having to worry about locking your data within your vendor cloud
    - Using Databricks' marketplace applications with DataBahn forking out data streams to different tables within Databricks
- **Get visibility into the health of telemetry generation**
    - By using the dynamic device inventory generated by DataBahn to keep track of devices to identify unexpected silences, log outages, and detecting any other upstream telemetry blind spots
- **Reduce overall costs of operating Databricks**
    - Removing the need for any staging locations or custom integrations by taking advantage of DataBahn's native streaming integration to load data directly into tables
    - By routing less-frequently accessed data sets and keeping a copy of your logs using Data Highway to low-cost storage infrastructure such as your cloud storage (S3 / Blob / GCP storage) while adhering to the same data models and using Databricks external tables to access them
    - By adopting the use of open data formats like Iceberg and storing data older than your standard retention periods outside of Databricks and using Iceberg tables to access them

Benefits of using DataBahn

Out-of-the-box connectors and integrations

DataBahn offers effortless integration and plug-and-play connectivity with a wide array of products and devices, allowing SOCs to swiftly adapt to new data sources.

Enrichment against Multiple Contexts

Enrich data against various contexts including Threat Intelligence, User, Asset, and Geo-location, providing a contextualised view of the data for precise threat identification.

Format Conversion and Schema Monitoring

The platform supports seamless conversion into any data model of your choice, facilitating flexible and faster downstream onboarding in Databricks.

Resilient data collection

DataBahn's highly resilient Smart Edge ensures that your team doesn't have to worry about single points of failures or managing occasional data volume bursts - the data collection never stops.

Reduced Costs

DataBahn helps you selectively extract key metadata based on frequency of usage, convert logs into insights to maximize retention of useful data whilst keeping costs of operating the warehouses optimal.

Risk-free data sharing

Use DataBahn to fork out data streams to different tables within Databricks for restricted data sharing to Databricks marketplace applications.

Get your data AI-ready

DataBahn's AI-ready framework gets your data cleansed, enriched, feature extracted, and with embeddings generated to build AI-powered apps.

Schema drift

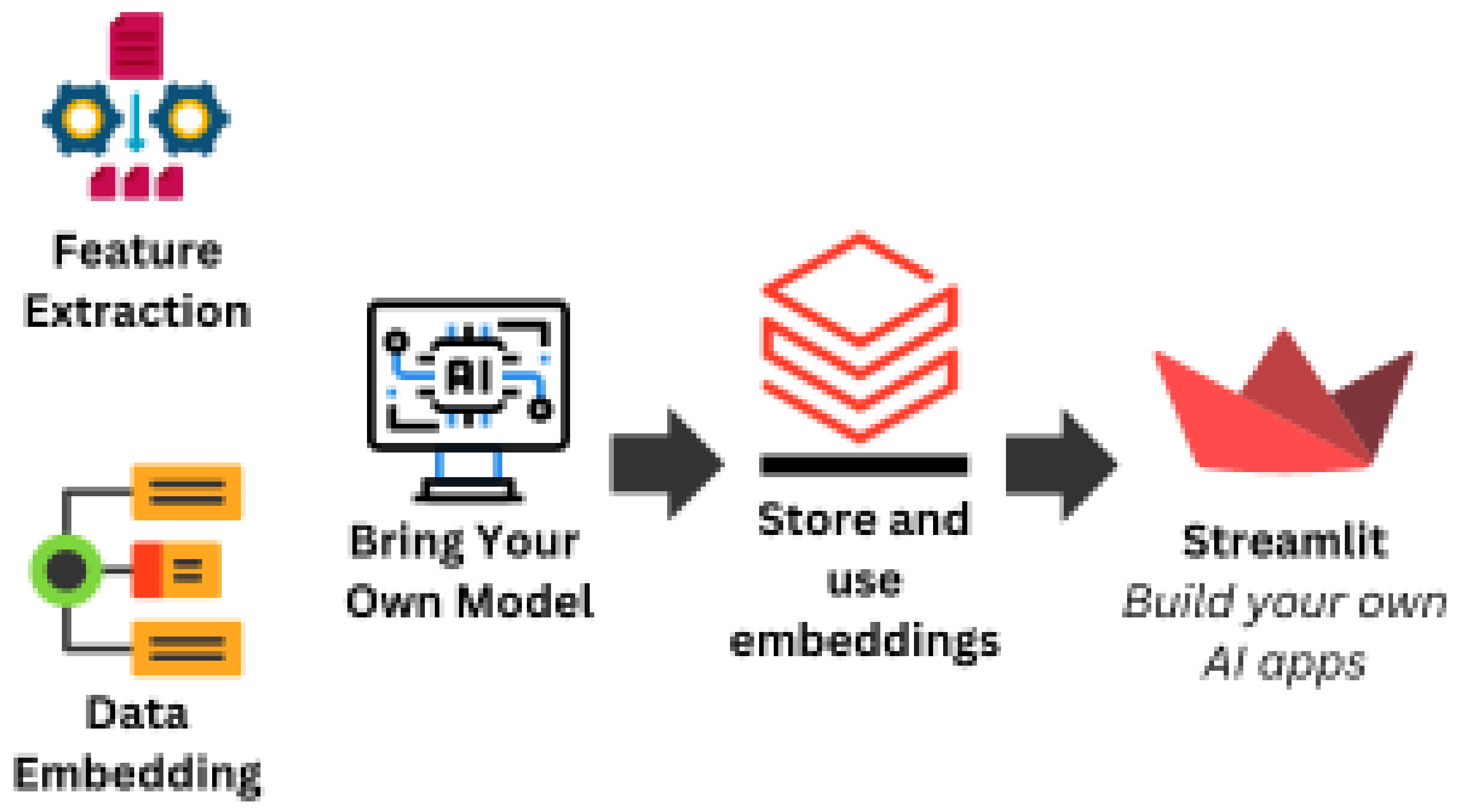
Detect changes to log schema intelligently for proactive adaptability and to avoid downstream detection impact.

Sensitive Data Detection

Identify, isolate, and mask sensitive data to ensure data security, governance, and compliance.

Relevance-based data orchestration

Tier and segment data based on relevance and move into different repos and tables so you can put purpose to your data.



Flexibility to your data stores

Deliver data to any external or iceberg tables and use Databricks to centrally query and access the data.

With DataBahn and Databricks, unlock the power of your data by maximizing the value while reducing the overhead it takes to collect and ingest data and the overall operating costs. DataBahn's purpose-built data collection and orchestration platform enables your teams to worry less about data acquisiting into Databricks.

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