CONFIDENTIAL

superna

Securing data pipelines at the storage layer - From SQL to Files/Objects

Andrew MacKay CTO & CSO

About Superna

- Over 5 EB data across 3600+ global customers
- HQ in Ottawa, Ontario and Boston with 110 global employees
- Profitable, cash-flow positive and investing for the future
- Profit 500 - Canada’s Fastest Growing Companies for 5 consecutive years
- Founded in 2008 to redefine unstructured data solutions

5+ EB
data under management

3600+
customers globally

8+
global locations

PROFIT 500
CANADA’S FASTEST-GROWING COMPANIES

Session Summary

AI/ML data pipelines consume data from file systems and object stores and structured databases using Trino to provide a data analytics platform.

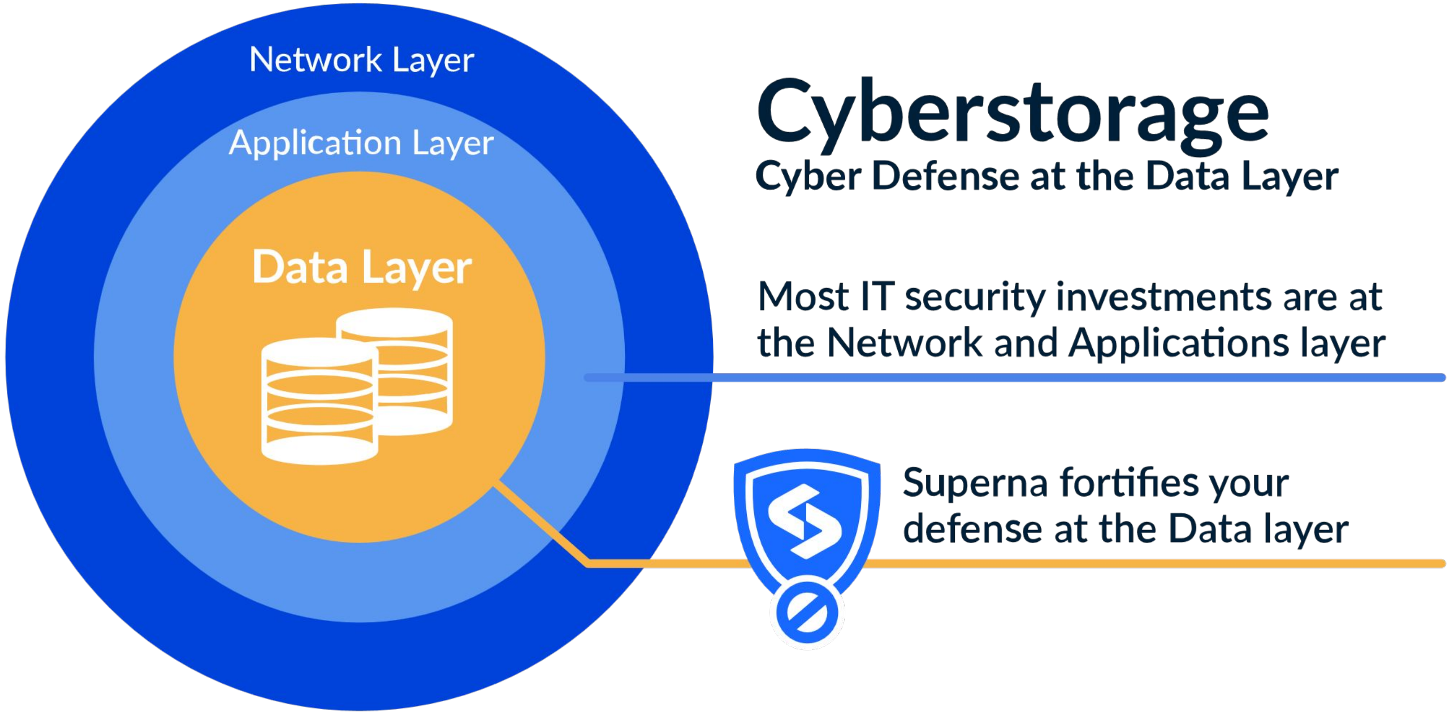
The Data Lake is the “weak link” in the AI/ML pipeline security posture

Learn how Superna protects your Data Lake including SQL security within Trino combined with storage layer security for file and object data stores

What is CyberStorage?

> “CyberStorage offers an *active* defense of storage systems and their data against cyber attacks through prevention, early detection and blocking of attacks, and aids in recovery through analytics and storage-specific recovery capabilities.”

Data Centric Security Framework



Cyberstorage

Cyber Defense at the Data Layer

Most IT security investments are at the Network and Applications layer

Superna fortifies your defense at the Data layer

The last line of defense – ***THE DATA!***

Prevention is the NEW Detection

Data Lakes & Security

Problem Statement

1. Combining data from structured and unstructured data sources to build a data lake creates a new “Attack Surface”
2. Separate File, Object and SQL security fragments capabilities to get a complete view

The Solution

1. Enable end to end chain of custody from SQL to the underlying files and objects that make up the Data Lake to address the security gap
2. Secure File, Object and SQL data manipulation with AI anomaly detection
3. Create a unified security layer for Data Lakes that monitors all data source DML activity and all data stores

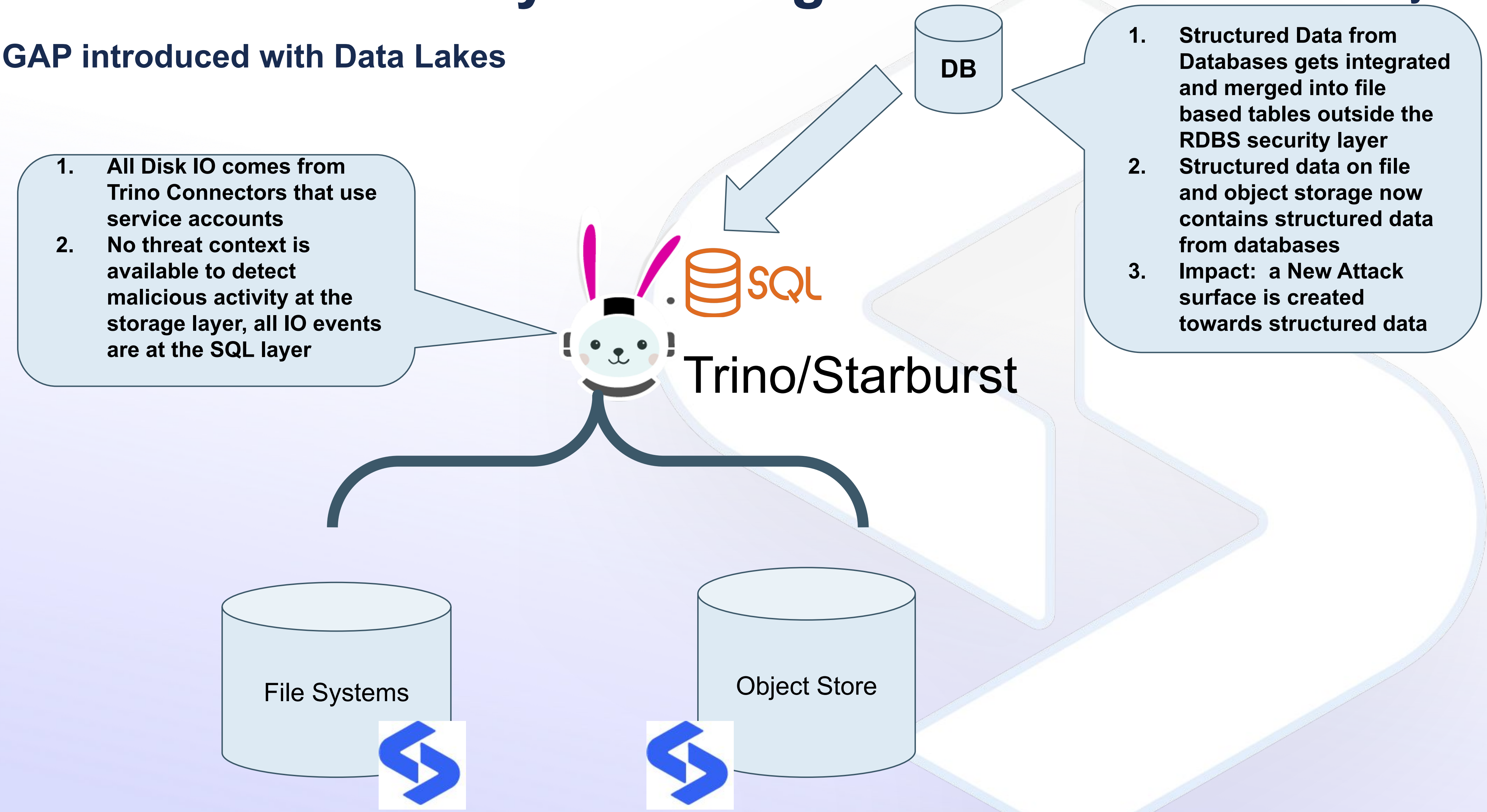
Protecting your AI model training source

> "30% of enterprises using AI reported having had a security or privacy breach against their AI environment."

Gartner®

Next Generation CyberStorage for Data Lakes Security

GAP introduced with Data Lakes



1. All Disk IO comes from Trino Connectors that use service accounts
2. No threat context is available to detect malicious activity at the storage layer, all IO events are at the SQL layer

1. Structured Data from Databases gets integrated and merged into file based tables outside the RDBS security layer
2. Structured data on file and object storage now contains structured data from databases
3. Impact: a New Attack surface is created towards structured data

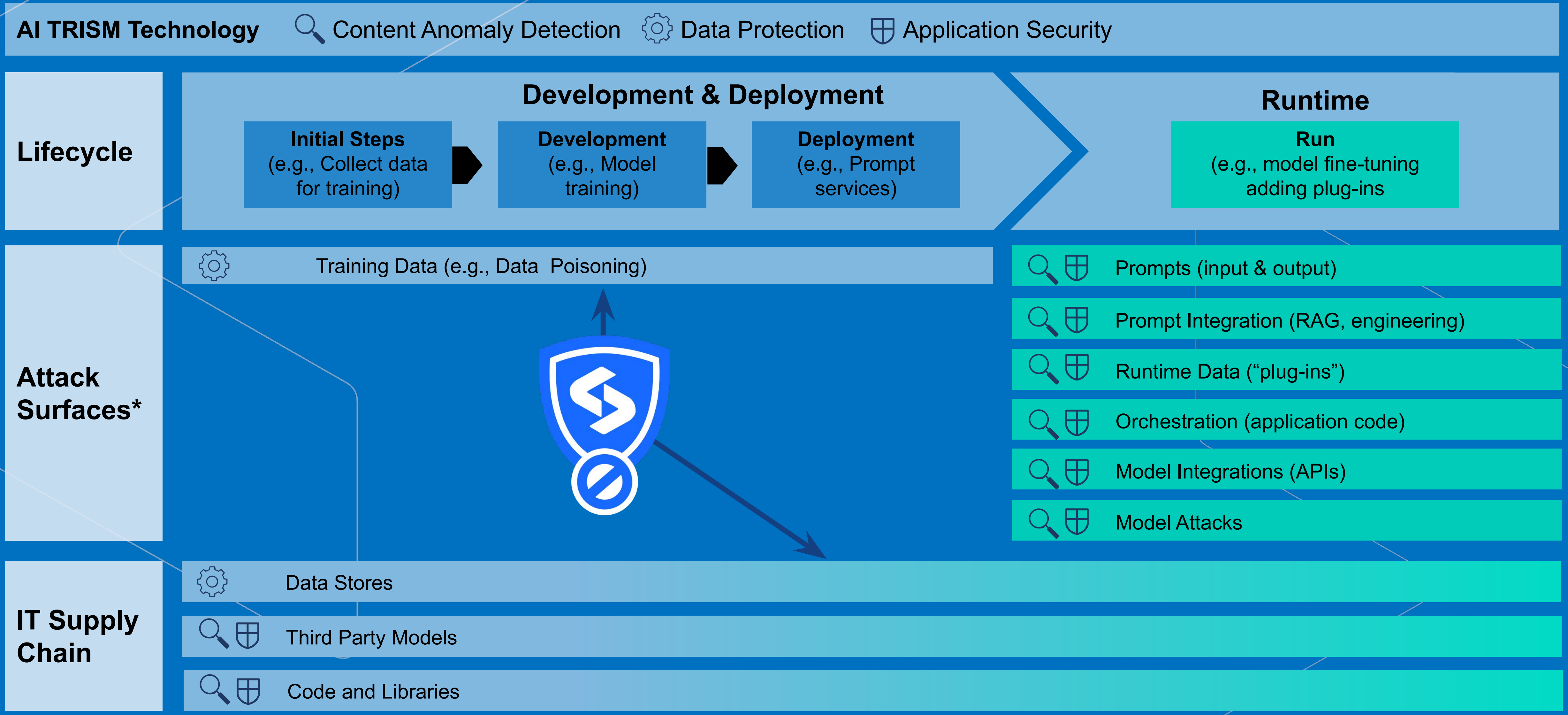
Achieving transparent security and integrity for your AI models

- Training data is the “weak link” in the AI/ML pipeline
- Each stage has vulnerabilities that impact integrity, traceability, resilience, and security

Stages of a Machine Learning Data Pipeline

- Data Collection
  - Data Cleaning and Preprocessing
  - Data Exploration and Analysis
  - Feature Engineering
  - Data Splitting
- Model Training
- Model Evaluation
- Model Tuning and Optimization
- Model Deployment
- Model Monitoring

AI Attack Surfaces



AI TRISM Technology 🔍 Content Anomaly Detection ⚙ Data Protection 🛡 Application Security

Lifecycle

Development & Deployment

Initial Steps (e.g., Collect data for training) → **Development** (e.g., Model training) → **Deployment** (e.g., Prompt services)

Runtime

Run (e.g., model fine-tuning adding plug-ins

Attack Surfaces*

⚙ Training Data (e.g., Data Poisoning)

🔍 🛡 Prompts (input & output)

🔍 🛡 Prompt Integration (RAG, engineering)

🔍 🛡 Runtime Data (“plug-ins”)

🔍 🛡 Orchestration (application code)

🔍 🛡 Model Integrations (APIs)

🔍 🛡 Model Attacks

IT Supply Chain

⚙ Data Stores

🔍 🛡 Third Party Models

🔍 🛡 Code and Libraries

* Main sample attack vectors only; others not shown. Source: Gartner

superna®

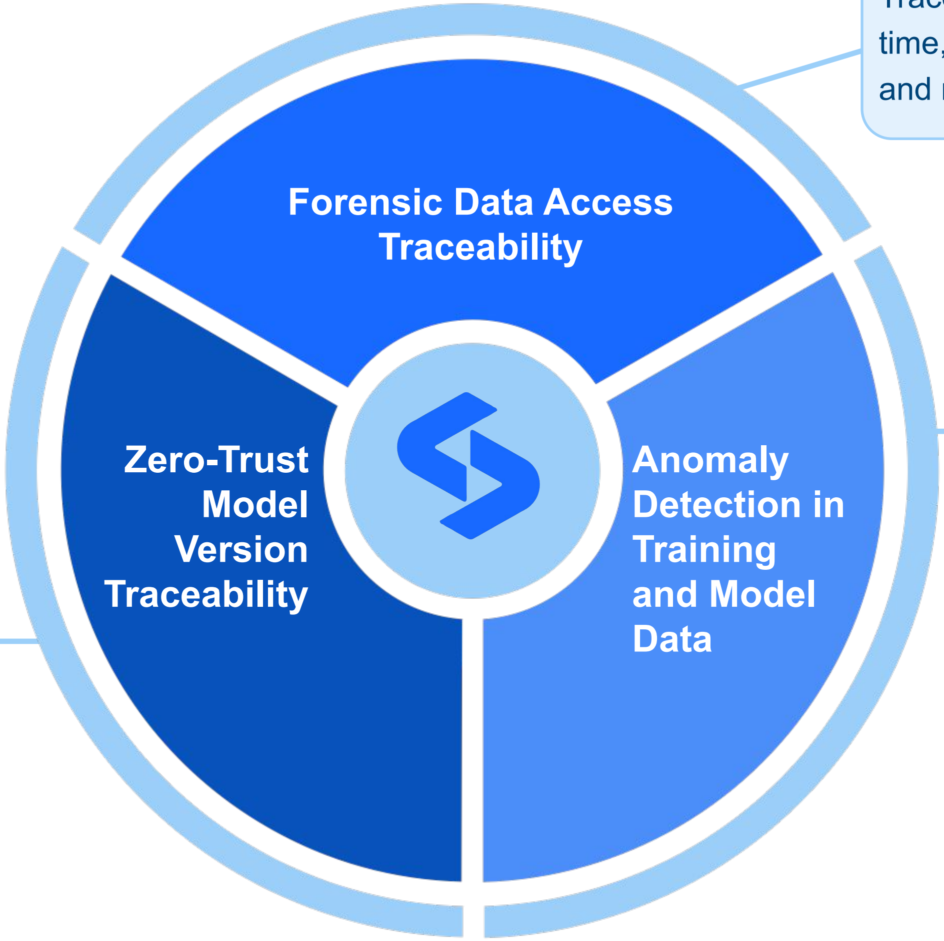
Superna’s Approach to Cyberstorage Security

Maintaining transparent security, trust and integrity of your AI models

Each model version and associated input data (raw input data, pre-processed data, split data) should be retained for long term traceability, stored as:

- An immutable copy
- Able to be recalled to any point in time in the past, for version integrity.

Monitor all access to versions of model data set, regardless of where it’s stored (file or object).


Forensic Data Access Traceability
Zero-Trust Model Version Traceability
Anomaly Detection in Training and Model Data

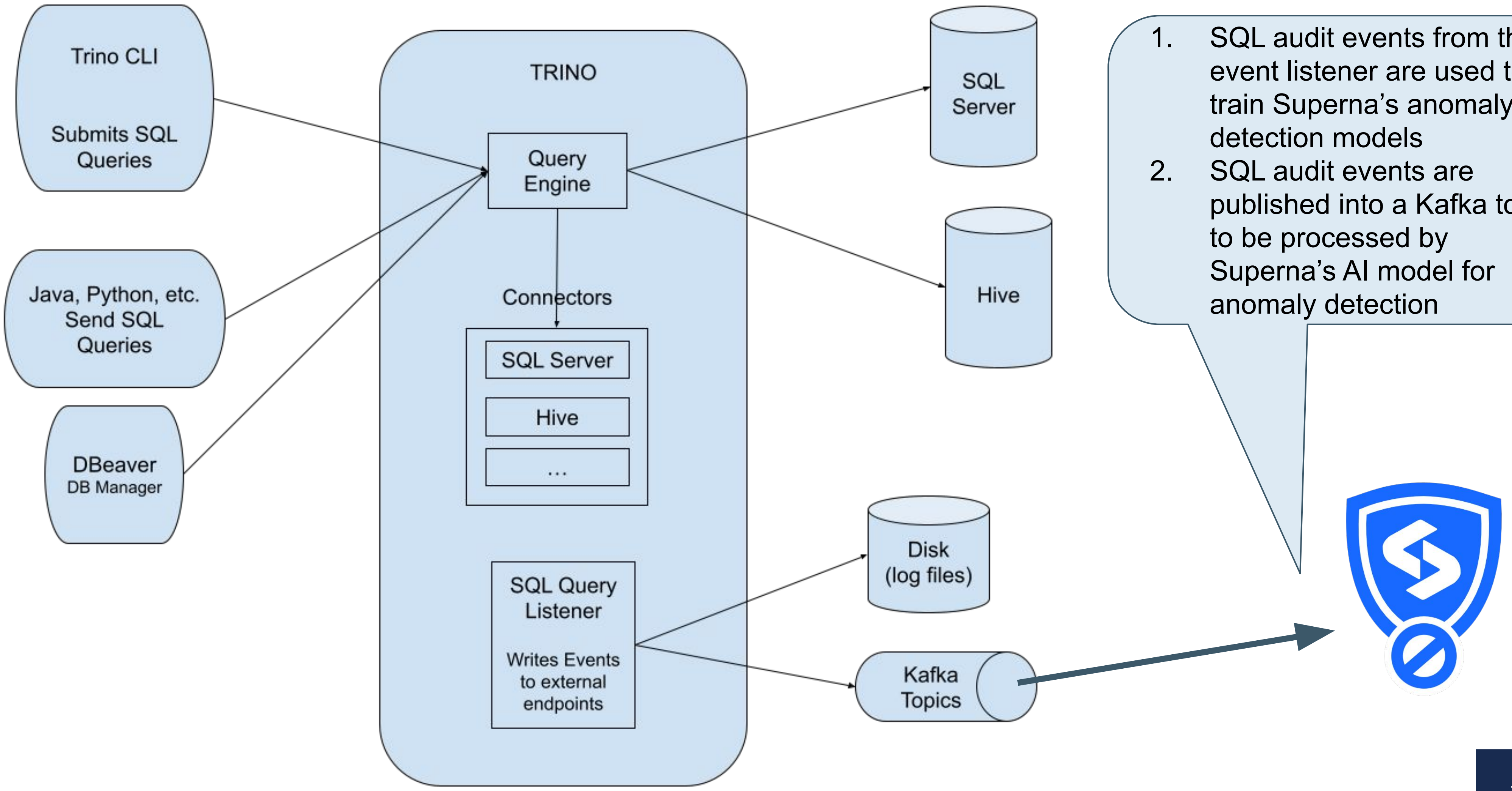
Trace data manipulation to a date, time, person or host, for training and model data

Detect changes to data deviating from **normal access patterns**:

- Malicious data modifications impacting model training
- Ransomware data attacks
- Unauthorized data deletion
- Model version attacks (e.g. replacing finished model with a different version
- Tampering w/ data cleaning/preprocessing outputs
- Model exfiltration

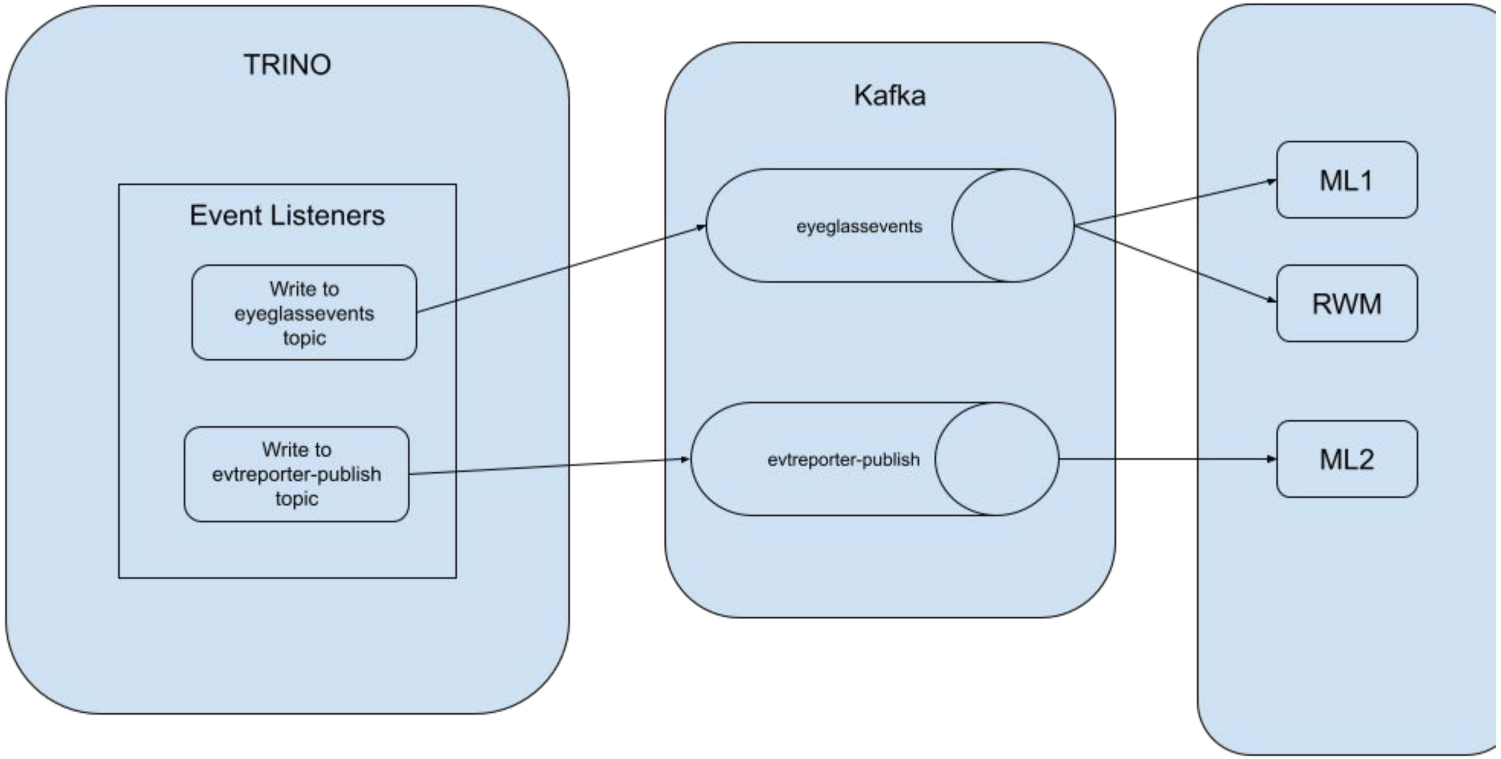
Trino SQL Security Demo

./trino http://trino-server-ip:8080



1. SQL audit events from the event listener are used to train Superna's anomaly detection models
2. SQL audit events are published into a Kafka topic to be processed by Superna's AI model for anomaly detection

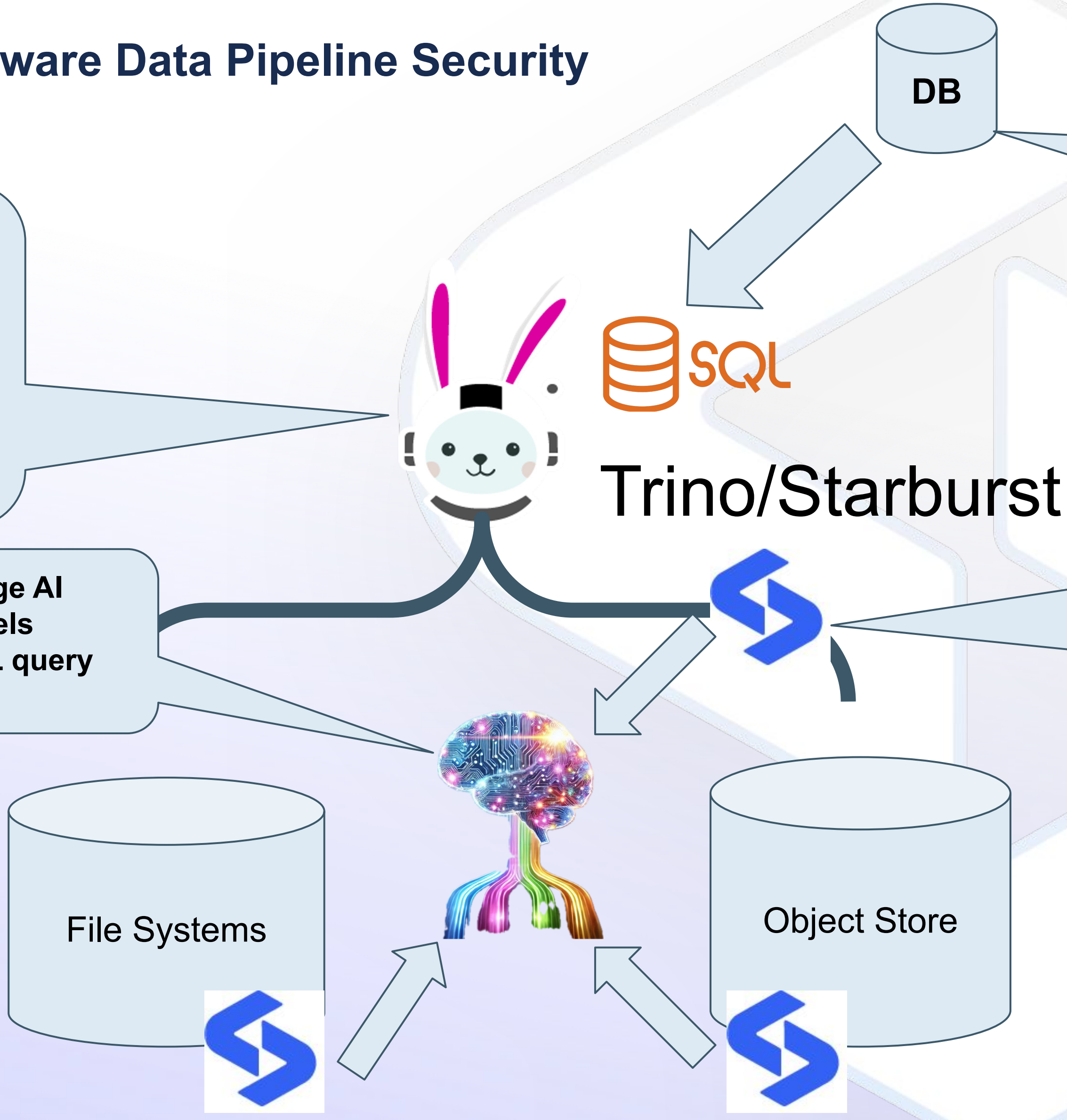
Superna Trino Event Listener Integrates with Suerna security edition



Next Generation CyberStorage for AI/ML Data Lakes

Context Aware Data Pipeline Security

1. Superna Trino Event plugin extracts user, sql command, source IP as threat feed into AI models
2. Combined with File/object activity to build a “End to End” Data threat AI anomaly detection capability



Traceability of structured data flowing into unstructured DB tables

Superna Event Listener Trino plugin extracts SQL Audit data and publishes to Kafka for processing by Superna ML models

CyberStorage AI Threat models include SQL query audit data

Trino Security Demo

1. A user Trino activity has been used to train a model on normal user activity
2. The anomaly detection model is looking for changes in behavior
3. Real time audit events from the Trino event listener are published to kafka for processing
4. A script is used to generate an anomaly user behavior on tables within Trino
5. The AI inference detects the anomaly

AI SQL Security Cyber Storage Anomaly Detection

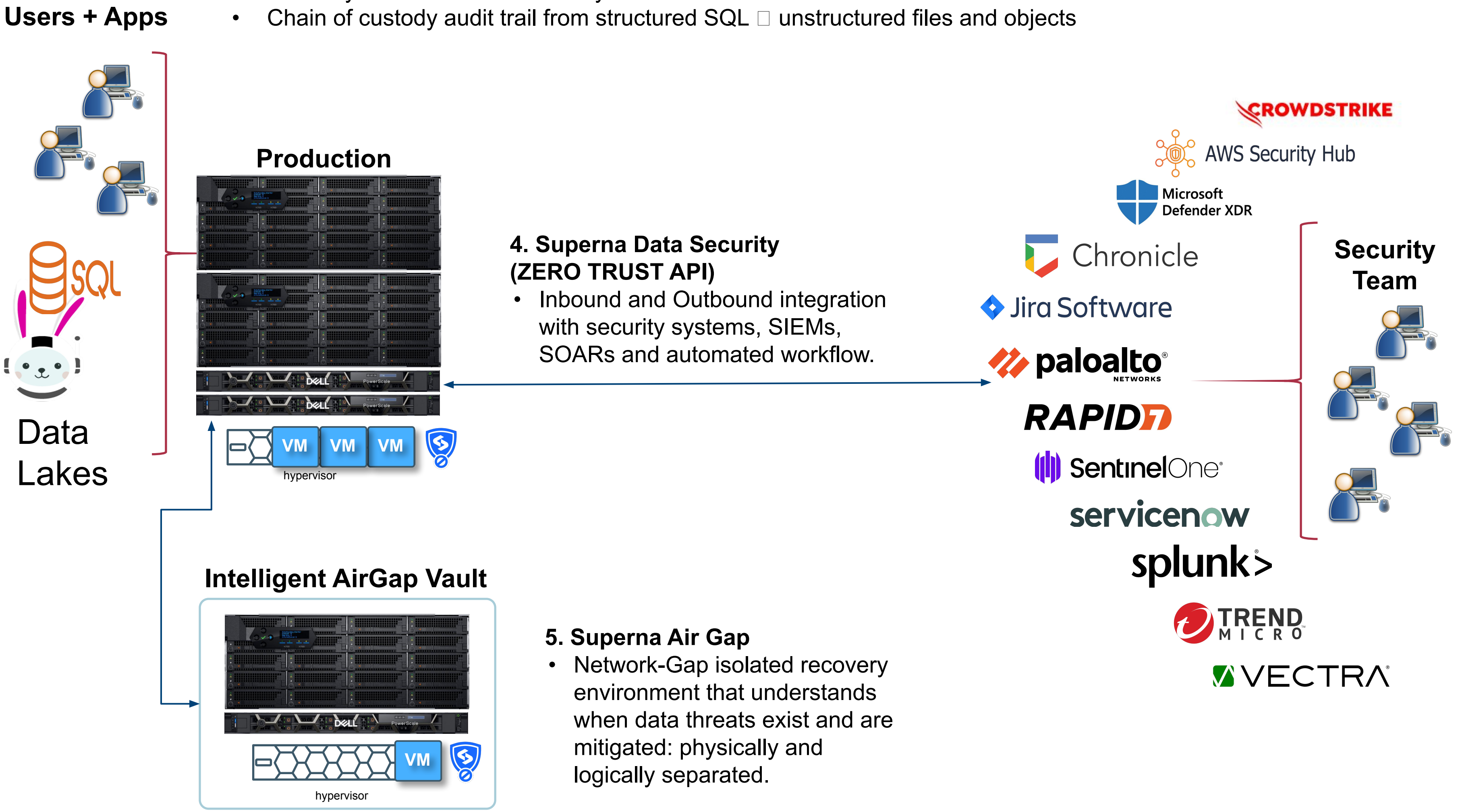
TRINO SQL SECURITY DEMO

Next Generation Cyberstorage Architecture for Data Lakes

1. **Unstructured & Structured Security for Data Lakes**
   - Superna developed plugin for Starburst sends audit data to Superna for analysis
   - Anomaly detection at the SQL layer with AI ML models
   - Chain of custody audit trail from structured SQL □ unstructured files and objects

2. Superna Data Security (DETECT)

- **SQL + NAS** Real time user behavior analysis (reads & writes). Intervention when abnormal deletes or writes on a per-user basis.
- Real-time Ransomware strain identification.
- Real-time Zero-Day attack identification.
- Real-time lockout of user, host or IP.



3. Superna Data Security (AUDIT & FORENSICS)

- **SQL data manipulation** audit history for Zero Trust analytics
- Applies historical behavior to real-time detection and alerting
- Geofencing and sensitive data share alerting
- Captures all forensics of an attack – host ID, IP, shares, path

4. Superna Data Security (ZERO TRUST API)

- Inbound and Outbound integration with security systems, SIEMs, SOARs and automated workflow.

5. Superna Air Gap

- Network-Gap isolated recovery environment that understands when data threats exist and are mitigated: physically and logically separated.