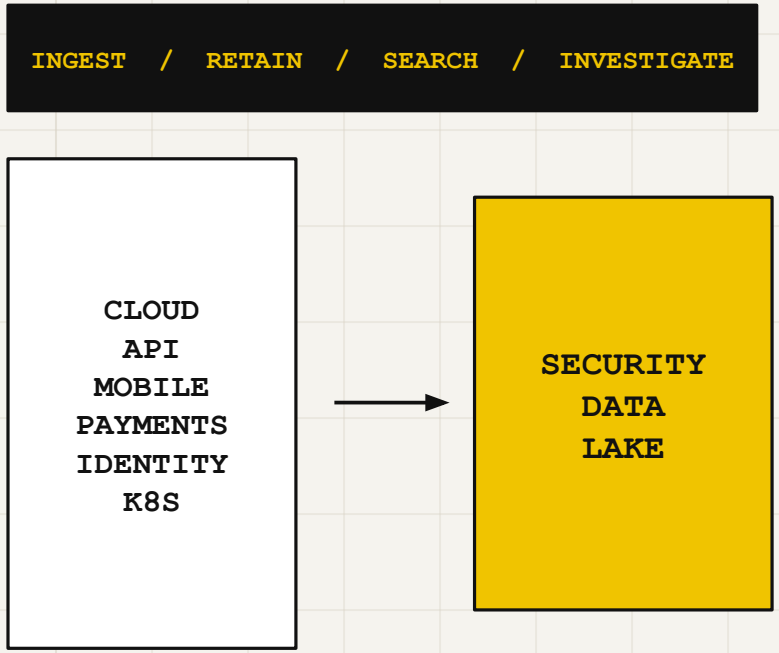
The "Security Data Lake"

vs.

The Legacy SIEM

ROVER
Agentic SIEM | SecOps Platform

SECURITY DATA LAKE VS LEGACY SIEM



Security data grew. SIEM architecture did not.

■ THESIS

ROVER
Agentic SIEM | SecOps Platform

SECURITY DATA GREW 100X

SIEM DIDN'T

10 YEARS AGO: FIREWALLS + ENDPOINTS

TODAY: CLOUD + APIS + SAAS + AI

QUESTION: WHAT IS THE DATA FOUNDATION?

The bottleneck moved from "collect logs" to "store, search, retain, and use logs economically."

AGENDA

ROVER
Agentic SIEM | SecOps Platform

WHAT WE WILL COVER

Start with SIEM basics, then expose the bottlenecks, then show how Security Data Lakes change the model.

- What is SIEM and how does it work?
- The six taxes: ingestion, retention, warehouse, format, compliance, search
- Product names and public pricing signals
- What is a Security Data Lake and how it solves the same problems
- Problems that Security Data Lakes introduce

PROMISE

This is not a “SIEM is dead” talk. It is a “SIEM needs a better data foundation” talk.

■ DEFINITION

ROVER
Agentic SIEM | SecOps Platform

WHAT IS SIEM?

Security Information and Event Management

- A central platform for security events and logs
- Collects data from devices, users, applications, cloud, and tools
- Correlates activity to detect suspicious behavior
- Creates alerts, dashboards, investigations, and compliance reports

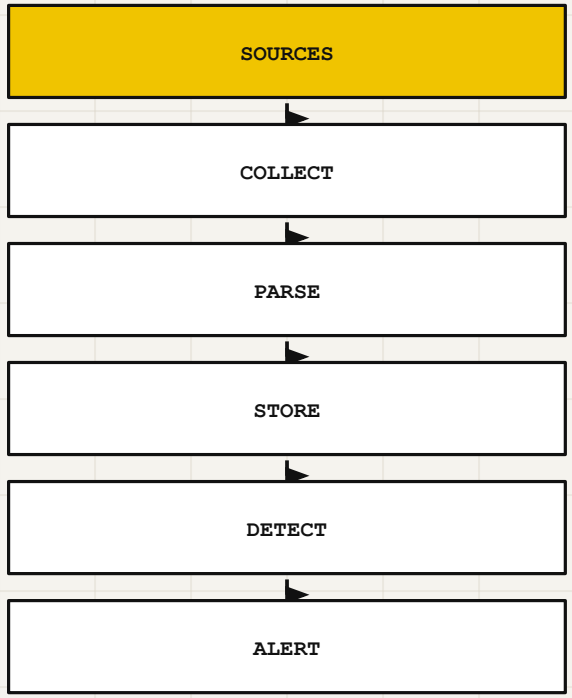
SIMPLE VERSION

A SIEM is the security control room. It helps answer: is something bad happening, where, and what should we do next?

HOW DOES SIEM WORK?

A SIEM is basically a security data pipeline plus an analyst workflow layer.

- Collect from many sources using agents, APIs, syslog, cloud connectors, or streams
- Parse raw logs into structured fields like user, IP, host, event type, and timestamp
- Normalize and index the data so analysts can search it
- Run correlation rules and detections to create alerts and cases



■ VALUE

ROVER
Agentic SIEM | SecOps Platform

WHAT SIEM IS GOOD AT

SIEM remains valuable. The problem is the warehouse role, not the SOC role.

- Real-time alerts and correlation rules
- Central visibility for the SOC
- Dashboards, reports, and compliance views
- Case management and investigation workflow
- Threat hunting across high-value telemetry

IMPORTANT

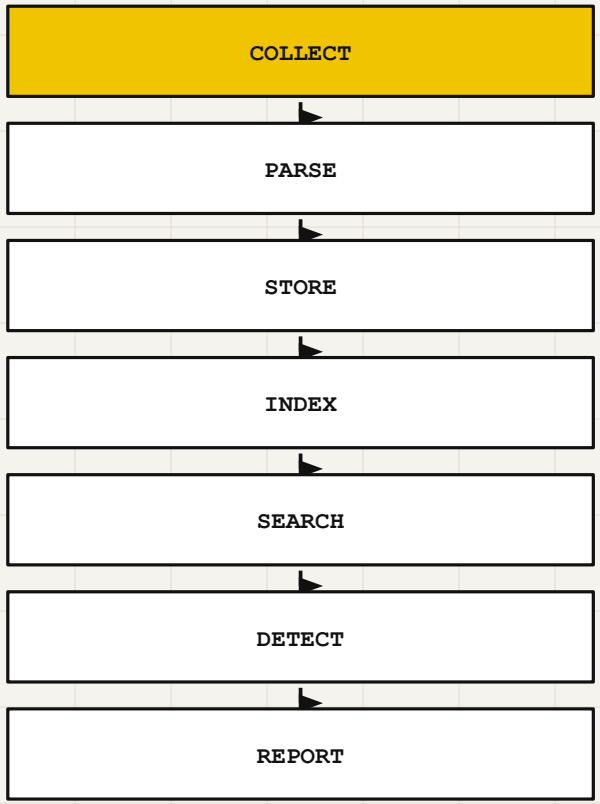
This talk is not saying SIEM is bad. It is saying SIEM should not always be the only warehouse for every byte of security data.

ROVER
Agentic SIEM | SecOps Platform

WHERE THE BOTTLENECK BEGINS

One platform is asked to do every job in the security data lifecycle.

- Collector: bring data in from many systems
- Parser: convert messy logs into fields
- Warehouse: store and retain all events
- Search engine: return answers during investigations
- Detection engine: run alerts and correlation rules
- Compliance tool: prove what happened months later

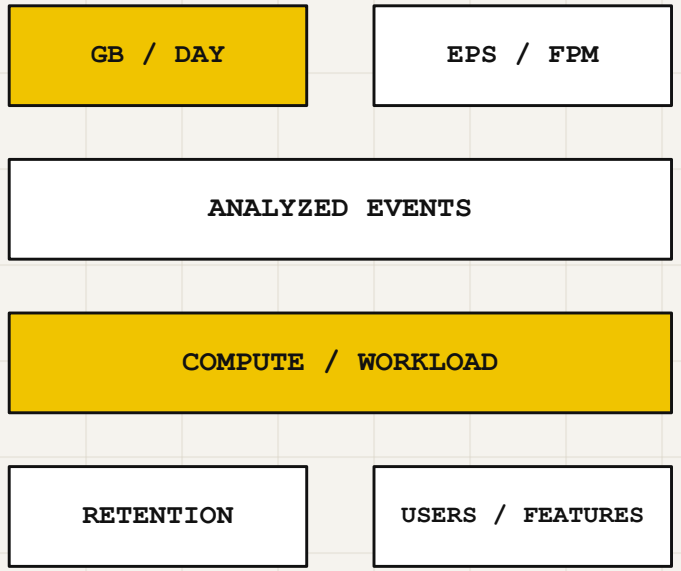


ROVER
Agentic SIEM | SecOps Platform

SIEM PRICING METERS

The pricing meter becomes the architecture pressure point.

- Some tools meter by GB/day ingested or retained
- Some meter by events per second or analyzed events
- Some meter by compute, workload, or scanned data
- Some package retention, users, SOAR, AI, and advanced features separately
- The result: every data-source decision can become a budget decision



Visibility should be driven by risk, not licensing.

ROVER
Agentic SIEM | SecOps Platform

SIEM PRODUCTS & PRICING SIGNALS

Public pricing signals show different meters: ingest, events, retention, compute, packages.

Product	Pricing signal	Presenter line
Splunk	Ingest, workload, entity, and activity-based pricing models; security pricing via estimate/sales.	Classic example of ingestion/workload economics. Data volume becomes a budget conversation.
Microsoft Sentinel	Analytics tier with PAYG/commitment; data lake tier meters include ingest, processing, query, storage.	Useful because Microsoft now explicitly separates analytics tier and data lake tier.
IBM QRadar SIEM	Usage model based on EPS and FPM; enterprise model based on managed virtual servers.	Good example of event-rate pricing rather than pure GB/day.
Datadog Cloud SIEM	$5 per 1M analyzed events/month billed annually; logs ingestion also listed separately.	Shows layered cost: ingest logs, index logs, then analyze security events.

ROVER
Agentic SIEM | SecOps Platform

MORE SIEM PRICING EXAMPLES

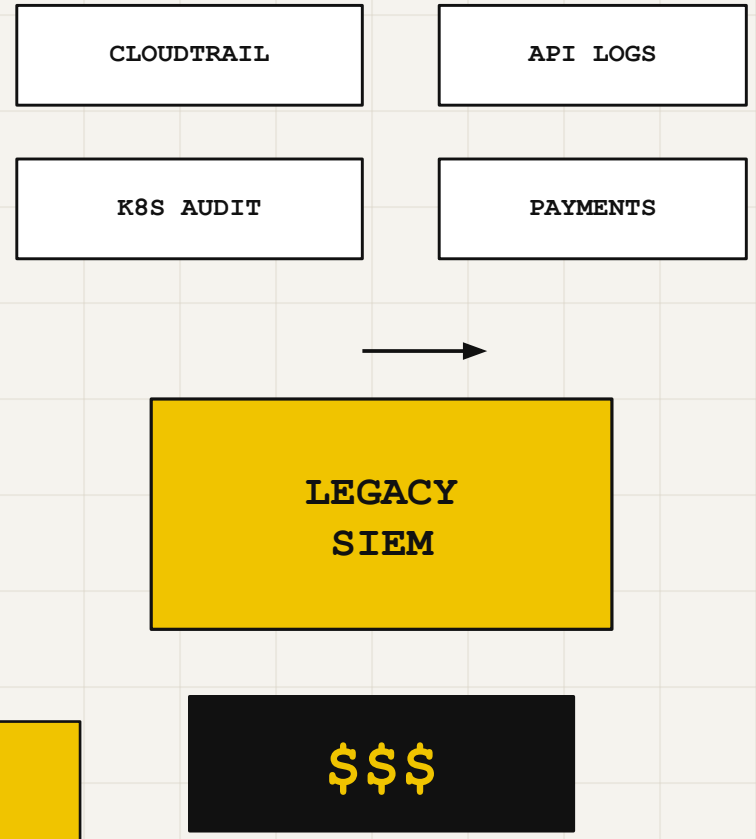
Newer SIEMs are already reacting to the same cost-pressure story.

Product	Pricing signal	Presenter line
Elastic Security	Serverless Security Analytics Essentials: as low as $0.09/GB ingest and $0.017/GB-month retention.	Transparent data-volume pricing; useful for modern SIEM cost discussion.
Sumo Logic Cloud SIEM	Enterprise Suite includes SIEM ingest variable, unlimited log capacity, customer-defined retention.	Shows packaged pricing with SIEM ingest as a variable.
Securonix	Moved to GB/day pricing and tiered packaging; promotes predictable capacity bands.	Good example of vendor response to pricing complexity.
Google SecOps	Package-based, ingestion-based; includes 12 months hot security telemetry retention; contact sales.	Strong contrast around long included retention and managed SecOps packaging.

■ TAX #1

ROVER
Agentic SIEM | SecOps Platform

THE INGESTION TAX



More logs = more cost

ROVER
Agentic SIEM | SecOps Platform

INGESTION BOTTLENECK

The SIEM is acting as collector, processor, warehouse, and index.

- Every event must pass through the SIEM pipeline
- Parsing CPU grows with every new data source
- Indexing workload grows with data volume
- License and infrastructure cost scale with visibility
- Security teams start negotiating with budget instead of risk

The bottleneck is not log generation. It is SIEM ingestion economics.

PROBLEM

ROVER
Agentic SIEM | SecOps Platform

WHY INGESTION TAX HURTS

It creates blind spots before the investigation even starts.

- Dropped logs remove evidence permanently
- Sampling can hide rare attacker behavior
- Excluded sources become unmonitored attack paths
- Verbose logs are often the first to be cut, but often useful in forensics
- The log you drop today may be the log you need in six months

BAD DECISION LOOP

More environment complexity → more logs → higher SIEM bill → fewer logs → less visibility.

■ TAX #2

ROVER
Agentic SIEM | SecOps Platform

THE RETENTION CLIFF

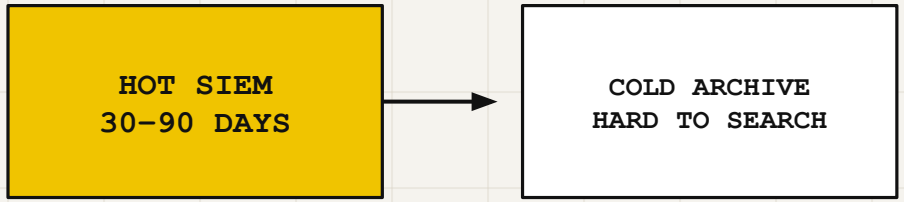
Attack found today. Logs expired yesterday.

ROVER
Agentic SIEM | SecOps Platform

RETENTION BOTTLENECK

The SIEM can store data, but long-term searchable retention is expensive.

- Hot indexed storage is not the same as cheap object storage
- Every retained event may need index, replication, and search capacity
- The cost curve pressures teams toward shorter retention
- Archive is often technically retained but operationally difficult to search
- Investigation history falls off a cliff



QUESTION:
WHEN DID THIS START?

PROBLEM

ROVER
Agentic SIEM | SecOps Platform

WHY RETENTION CLIFF HURTS

No history means no timeline, no scope, no proof, and no confidence.

- Credential theft may remain hidden for months
- Insider threats may take a year to discover
- Supply-chain compromises can be detected long after initial access
- Compliance may require one year or more of evidence
- Without old logs, the team cannot reconstruct the story

INCIDENT QUESTION

“When did this start?” is usually the first serious question. Retention decides whether you can answer it.

■ TAX #3

ROVER
Agentic SIEM | SecOps Platform

SIEM AS A LENS

NOT A WAREHOUSE

The warehouse stores the data. The SIEM provides the view.

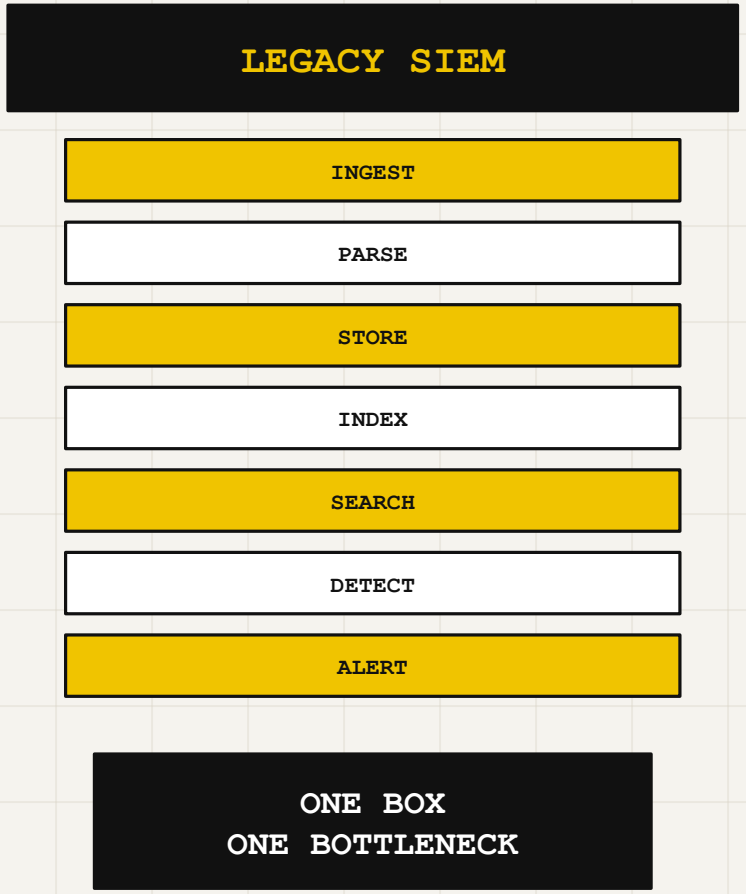
■ OLD MODEL

ROVER
Agentic SIEM | SecOps Platform

MONOLITHIC SIEM BOTTLENECK

Everything scales together. Nothing scales independently.

- Ingestion pressure affects detection coverage
- Storage pressure affects retention
- Search pressure affects analyst speed
- Parser pressure affects onboarding
- Vendor pressure affects the whole SOC



■ NEW MODEL

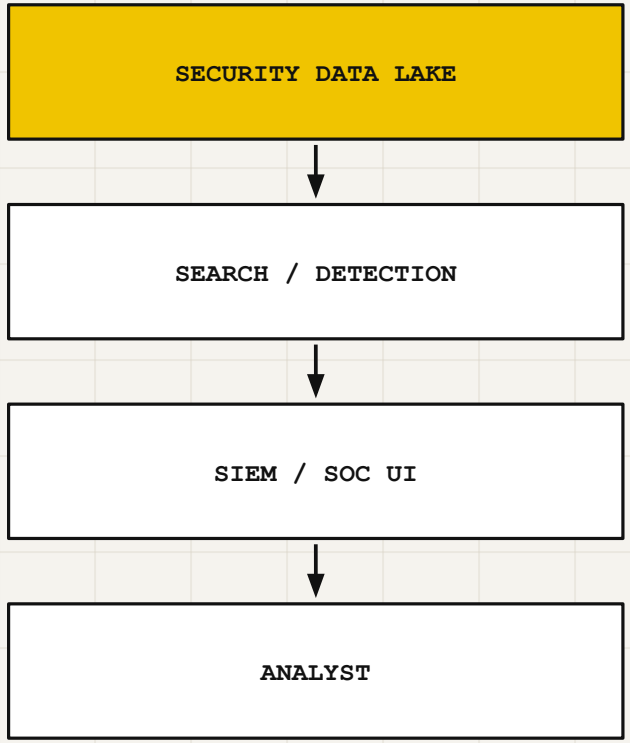
ROVER
Agentic SIEM | SecOps Platform

THE LENS MODEL

The data foundation and analyst experience are separated.

- Security Data Lake stores raw, normalized, and historical data

- Detection engines run on the right data at the right speed
- Search and analytics can scale separately from storage
- SIEM remains the analyst workflow and investigation lens
- Vendor lock-in is reduced because data is not trapped in one tool



SIEM does not disappear. Its job changes.

ROVER
Agentic SIEM | SecOps Platform

THE FORMAT TAX

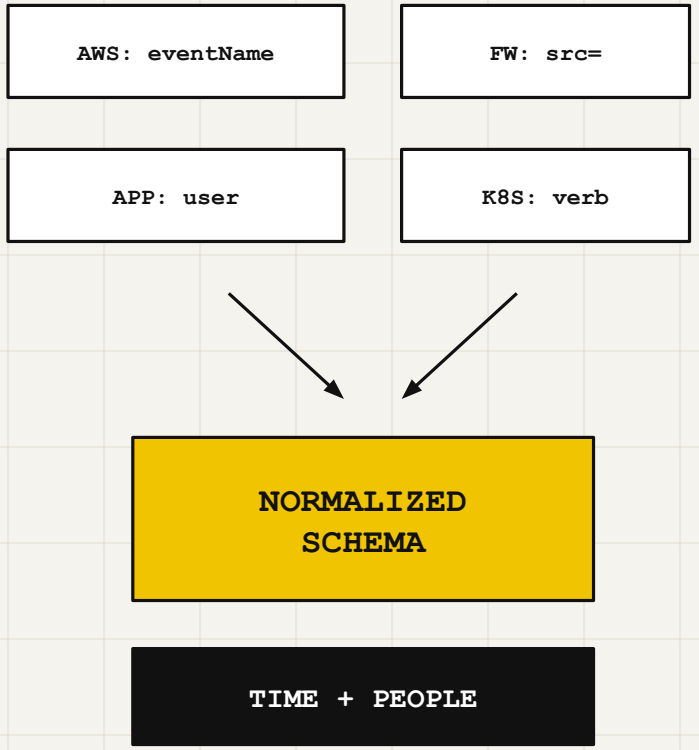
Every source speaks a different language.

ROVER
Agentic SIEM | SecOps Platform

WHY FORMAT TAX HURTS

Engineers become data janitors instead of threat hunters.

- New sources require parser development and validation
- Vendors change log formats and break mappings
- Field names differ: src_ip, source.ip, client_ip, remoteAddress
- Timestamps, user IDs, timezones, and asset names are inconsistent
- Human effort becomes the scaling bottleneck



Security teams should find threats, not translate logs.

ROVER
Agentic SIEM | SecOps Platform

THE COMPLIANCE TAX

Data exists. Evidence is scattered.

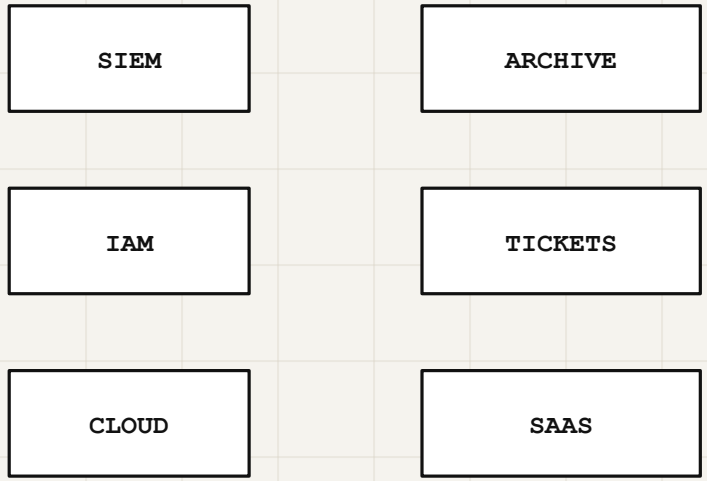
PROBLEM

ROVER
Agentic SIEM | SecOps Platform

WHY COMPLIANCE TAX HURTS

Audit becomes a manual data collection project.

- Evidence is spread across SIEM, archives, cloud, IAM, tickets, and SaaS tools
- Teams export CSVs, screenshots, reports, and logs by hand
- Manual stitching increases errors and delays
- Repeated audits recreate the same work every cycle
- Regulated systems need faster and more reliable proof



The best compliance strategy is good data architecture.

ROVER
Agentic SIEM | SecOps Platform

THE SEARCH TAX

Petabytes change everything.

■ PROBLEM

ROVER
Agentic SIEM | SecOps Platform

WHY SEARCH TAX HURTS

Waiting for a query is not incident response.

- Slow queries slow triage and containment
- Analysts narrow queries to avoid expensive searches
- Older data is skipped because it is too slow or too costly to search
- Investigations become biased toward what is already indexed
- Every delay gives the attacker more time

Waiting for a query is not incident response.

ANALYST LOOP

Search → wait → refine → wait → export → pivot → wait again.

PATTERN

ROVER
Agentic SIEM | SecOps Platform

THE COMMON THEME

These are not only security problems. They are data architecture problems.

- Ingestion Tax → data volume and routing problem
- Retention Cliff → storage tier and index economics problem
- SIEM Warehouse → tight coupling and lock-in problem
- Format Tax → schema and data engineering problem
- Compliance Tax → governance and fragmentation problem
- Search Tax → query compute and data layout problem

ROOT CAUSE

Security teams are hitting data-platform limits: cost, retention, schema, governance, and query scale.

■ DEFINITION

ROVER
Agentic SIEM | SecOps Platform

WHAT IS A SECURITY DATA LAKE?

A scalable repository for raw, normalized, enriched, and historical security telemetry.

- Stores security data independently from one SIEM hot tier
- Usually built on object storage or a cloud lakehouse
- Can keep raw logs and curated normalized tables together
- Lets multiple tools use the same data foundation
- Separates storage, compute, search, detection, and workflow

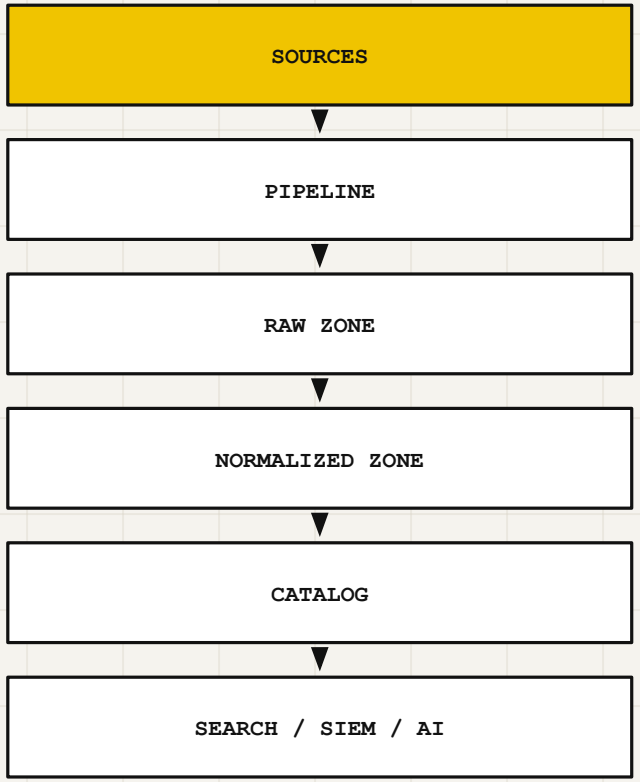
SIMPLE VERSION

The Security Data Lake is where security data lives. SIEM becomes one of the tools that uses it.

HOW SECURITY DATA LAKE WORKS

Raw first, normalize where valuable, govern everything, query with the right tool.

- Sources feed a telemetry pipeline
- Raw zone preserves original evidence
- Normalized zone supports consistent analytics and detections
- Catalog tracks schema, ownership, retention, and access
- SIEM, search, AI, notebooks, and compliance tools query the lake



ROVER
Agentic SIEM | SecOps Platform

DATA LAKE PRICING METERS

The Security Data Lake changes the cost model. It does not make data free.

- Storage: GB-month or TB-month
- Ingestion and normalization: GB processed
- Query scan: GB or TB analyzed
- Compute: warehouse, cluster, serverless, or security compute unit
- Operations: catalog, pipelines, egress, governance, indexing

STORAGE
QUERY SCAN
COMPUTE
NORMALIZATION
EGRESS
PIPELINES

COST MOVES
TO ARCHITECTURE

SECURITY DATA LAKE PRODUCTS & PRICING

Managed lakes and DIY lake architectures have very different cost meters.

Product	Pricing signal	Presenter line
Amazon Security Lake	Pay-as-you-go; AWS example: CloudTrail $0.75/GB, other AWS logs $0.25/GB, normalization $0.035/GB; S3 charges apply.	Managed security lake in your AWS account; strong for AWS-heavy orgs.
S3 + Athena DIY	S3 storage plus Athena query at $5/TB scanned; compression, partitioning, Parquet reduce scan cost.	Cheap storage, but poor data layout creates expensive searches.
Microsoft Sentinel Data Lake	$0.05/GB ingest, $0.10/GB processing, $0.026/GB-month storage, $0.005/GB query.	Good example of SIEM vendor adopting lake tier economics.
Cribl Lake	Cribl Lake public page shows 0.05 credits/GB; 1 credit = $1; pricing based on stored data.	Useful for telemetry routing + lake storage architecture.

MORE SECURITY DATA LAKE EXAMPLES

Some products are data-lake-native; others are SIEM experiences on top of lake storage.

Product	Pricing signal	Presenter line
Snowflake	On-demand storage listed at $23/TB/month; compute billed through warehouse credits and runtime.	Good for organizations already standardizing on Snowflake.
Databricks Lakewatch	Private preview; Databricks claims petabyte-scale investigations and up to 80% lower TCO.	Useful for lakehouse-first organizations and AI-native SecOps story.
Panther	Security Data Lake Search built on Snowflake; private/custom pricing via marketplace or sales.	Example of detection/search workflow on top of a security data lake.
Rover	S3-native search with S3-hosted indices; public site did not show list pricing.	Matches the theme: SIEM experience with S3-backed memory and evidence-cited cases.

■ SOLVE #1

ROVER
Agentic SIEM | SecOps Platform

SOLVING THE INGESTION TAX

Collect broadly. Analyze selectively.

LEGACY SIEM	DATA LAKE MODEL
• More logs • Higher SIEM bill • Drop / sample logs • Blind spots	• Collect broadly • Route intelligently • Analyze selectively • Retain evidence

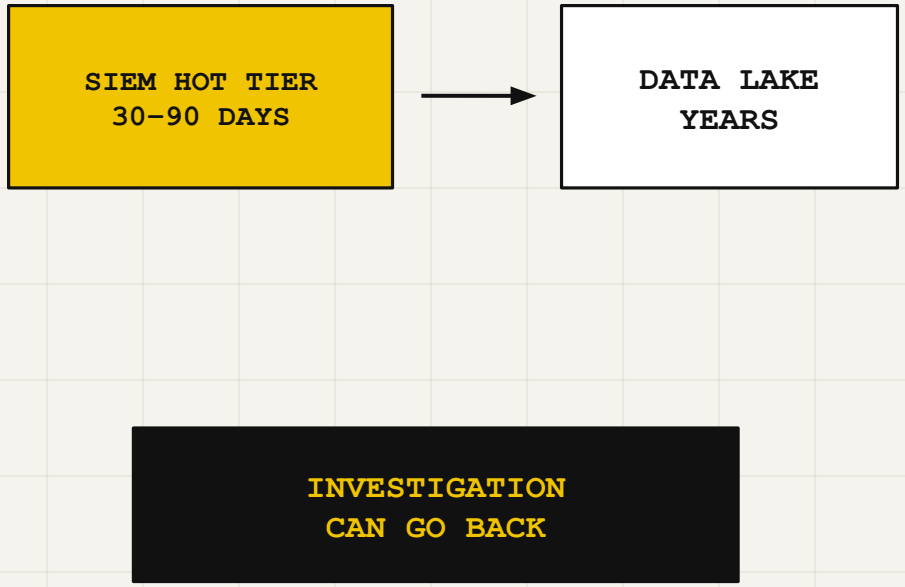
■ SOLVE #2

ROVER
Agentic SIEM | SecOps Platform

SOLVING THE RETENTION CLIFF

Years of data, not just 90 days.

- Long-term storage is separated from hot SIEM indexing
- Old data remains available for forensics and compliance
- Search/query can be run when needed using scalable compute
- Retention policy can follow risk, regulation, and investigation needs
- Historical context becomes possible for AI and threat hunting



Retention follows risk and regulation, not hot-index economics.

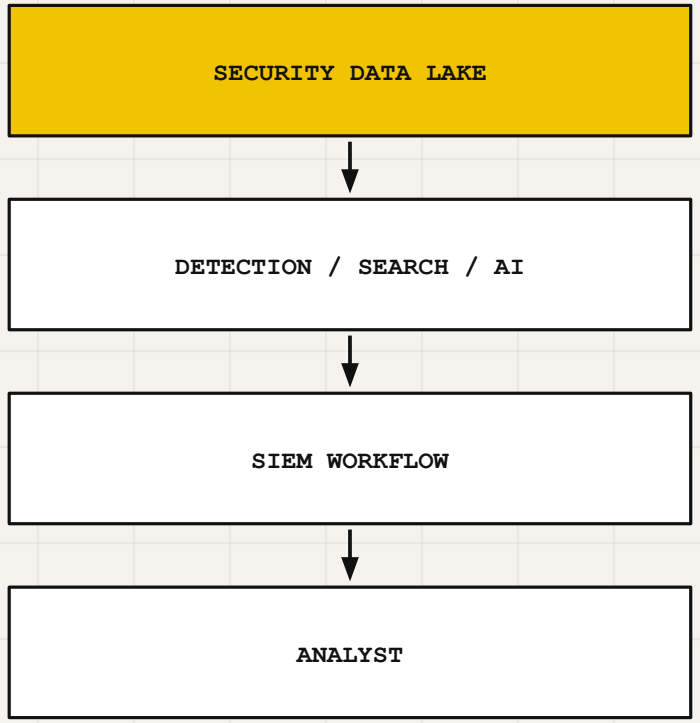
■ SOLVE #3

ROVER
Agentic SIEM | SecOps Platform

SOLVING SIEM-AS-WAREHOUSE

Lake becomes the memory. SIEM becomes the lens.

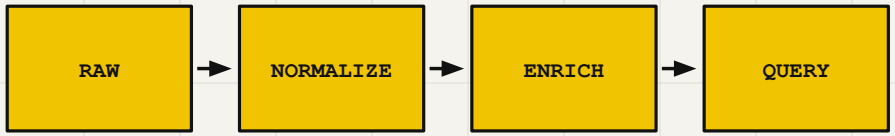
- Data remains in an open, scalable foundation
- SIEM is one consumer, not the owner of all data
- Search, detection, AI, compliance, and notebooks can share the same base
- Vendor migration becomes less painful
- Storage, compute, detection, and workflow scale independently



SOLVING THE FORMAT TAX

Raw first. Normalize when needed. Standardize over time.

- Raw data can be stored immediately without blocking on perfect parsers
- Important sources can be normalized first
- Multiple schemas can coexist: raw, OCSF, ECS, UDM, internal schema
- Columnar formats such as Parquet reduce query cost and improve performance
- Schema work becomes incremental instead of a gate before collection



■ SOLVE #5

ROVER
Agentic SIEM | SecOps Platform

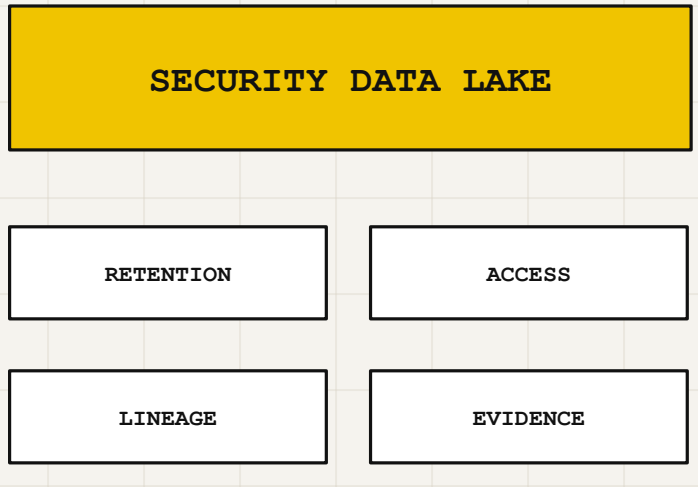
SOLVING THE COMPLIANCE TAX

One governed place for security evidence.

- Central retention policy instead of scattered evidence stores

- Raw records preserved for investigation and audit trails
- Governed access with encryption, logging, and role controls

- Repeatable audit queries instead of manual exports
- Compliance becomes a byproduct of operational logging

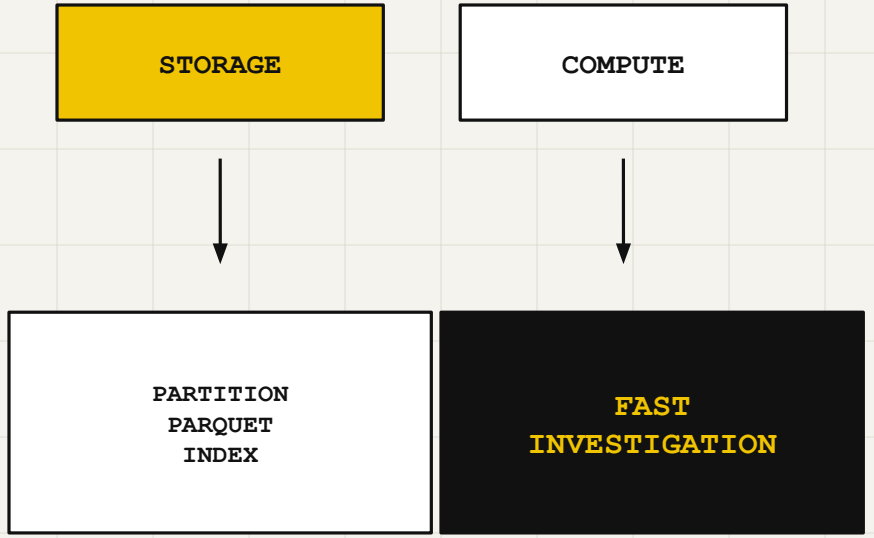


ROVER
Agentic SIEM | SecOps Platform

SOLVING THE SEARCH TAX

Scale compute separately from storage.

- Use real-time detection for urgent signals
- Use lake query for historical investigation
- Use distributed compute for high-volume hunting
- Use precomputed summaries for repeated questions
- Use S3/Snowflake/Databricks/Sentinel/other search layers depending on workflow



A lake needs search architecture, not just storage.

■ PROBLEM #1

ROVER
Agentic SIEM | SecOps Platform

COST MOVES. IT DOES NOT DISAPPEAR.

A data lake lowers cost only if the data layout is intelligent.

- Storage may be cheap, but query scans and compute can surprise you
- Normalization, pipelines, catalog, indexing, and egress add cost
- Poor partitioning can make every investigation scan too much data
- Duplicate storage across SIEM, lake, archive, and observability tools wastes money
- FinOps must be part of security data architecture

NEW COSTS

Storage
Query scans
Compute
Pipelines
Catalog
Egress
Indexes
Operations

■ PROBLEM #2

ROVER
Agentic SIEM | SecOps Platform

DETECTION LATENCY

Data Lake does not automatically mean real time.

- Some detections need seconds or minutes, not batch queries
- Ransomware, privilege escalation, payment fraud, and cloud takeover need fast response
- Streaming detection, EDR, NDR, cloud-native alerts, or SIEM analytics still matter
- Lake is excellent for context and history, but not always the first alert path
- The winning architecture is hybrid

RULE OF THUMB

Real-time signal → SIEM / detection engine
Long-term history → Security Data Lake
Investigation context → both

■ PROBLEM #3

ROVER
Agentic SIEM | SecOps Platform

ANALYST EXPERIENCE

SQL is not a SOC workflow.

- SOC analysts need pivots, timelines, alerts, cases, and evidence exports
- Raw tables alone are not enough for incident response
- If the lake is painful to use, analysts will avoid it
- A SIEM, search UI, or agentic investigation layer is still needed
- The lake is the foundation, not the analyst experience by itself

ANALYST NEEDS

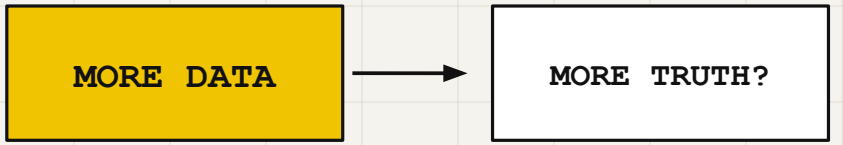
User timeline
Host timeline
IP pivots
Case notes
Evidence export
Alert context
Saved hunts

DATA QUALITY

Bad data at scale is still bad data.

- Wrong timestamps, missing users, broken JSON, local timezone problems
- Duplicate logs, dropped fields, inconsistent asset names
- No owner, environment, severity, or business-context mapping

- More data does not automatically mean more truth
- Source health, schema validation, and entity enrichment are required



■ PROBLEM #5

ROVER
Agentic SIEM | SecOps Platform

SECURITY & PRIVACY RISK

Central data becomes central risk.

- The lake may contain employee behavior, customer identifiers, admin actions, and authentication history
- Attackers can learn how the environment works if they access the lake
- The lake must be protected like a crown-jewel system
- Controls: encryption, least privilege, RBAC/ABAC, masking, audit logs, immutability
- Data residency and retention rules must be enforced

CROWN JEWEL

The security data lake must not become the easiest place to steal security data.

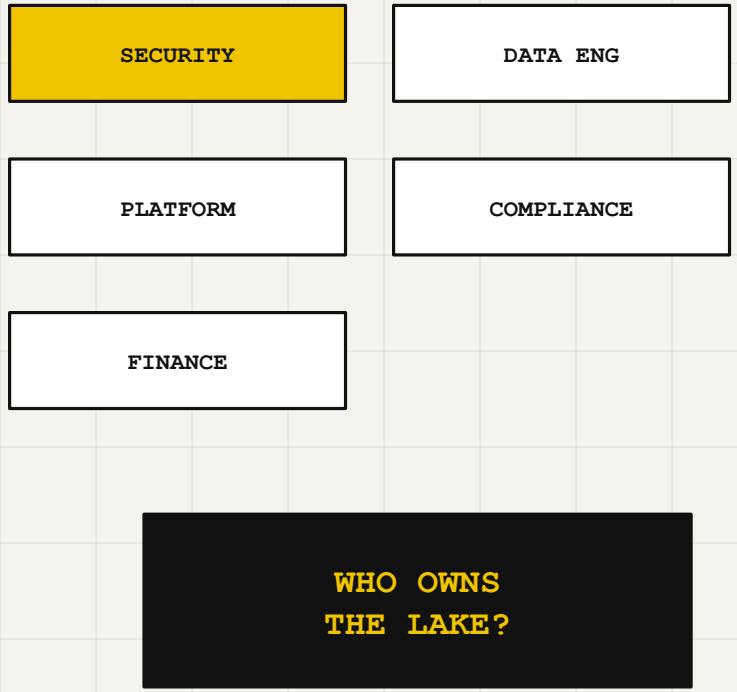
■ PROBLEM #6

ROVER
Agentic SIEM | SecOps Platform

OWNERSHIP & OPERATING MODEL

A Security Data Lake sits between teams. That is dangerous if nobody owns it.

- Security wants detection and investigation value
- Data engineering owns pipeline quality and schemas
- Cloud/platform teams own storage, IAM, and infrastructure
- Compliance owns retention and evidence requirements
- Finance wants cost controls and predictable usage
- Without clear ownership, the lake becomes nobody's product



HYBRID

ROVER
Agentic SIEM | SecOps Platform

THE WINNING MODEL IS HYBRID

SIEM for real-time workflow. Data Lake for scale and memory.

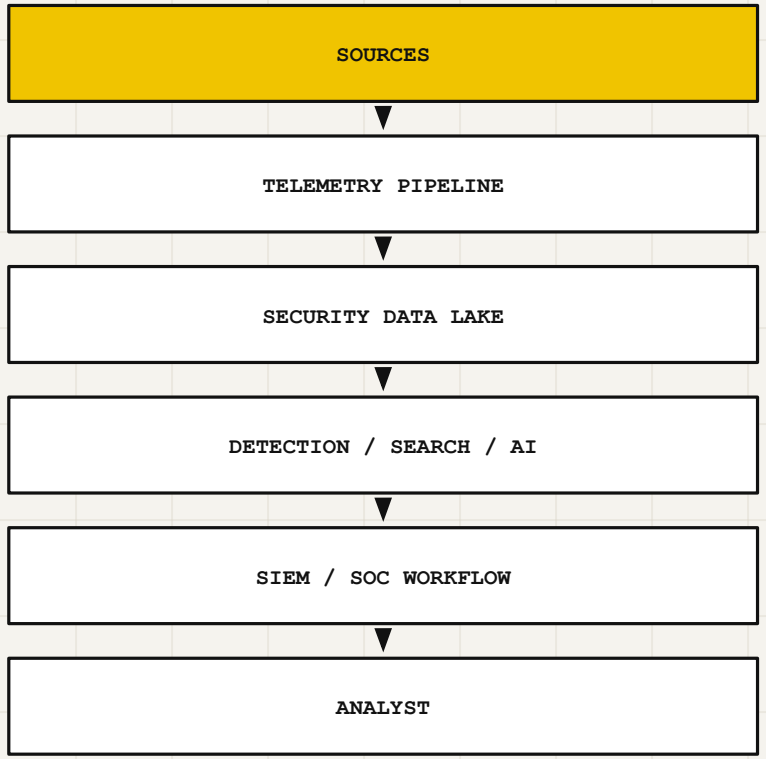
SIEM	DATA LAKE
• Real-time detections • Dashboards • Cases • SOC workflow • Compliance views	• Long retention • High-volume logs • Historical search • AI / ML analytics • Forensics evidence

ROVER
Agentic SIEM | SecOps Platform

TARGET ARCHITECTURE

Security data platform first. SIEM experience on top.

- Sources generate telemetry across cloud, identity, endpoint, app, network, and business systems
- Pipeline filters, enriches, redacts, transforms, and routes data
- Security Data Lake stores raw and normalized data with governance
- Detection, search, AI, and compliance tools operate on the foundation
- SIEM remains the SOC interface for alerts and response



■ DECISION

ROVER
Agentic SIEM | SecOps Platform

HOW TO DECIDE WHERE DATA LIVES

Do not ask only: can we afford to ingest this? Ask: what job does this data need to do?

- Need real-time alerting? Send to SIEM / detection tier
- Need long-term evidence? Store in the lake
- Need frequent investigation? Build summaries or indexes
- Need compliance proof? Govern retention and access
- Need ML/AI context? Normalize and enrich entity views

NEW QUESTION

Where should this data live?
How fast do we need it?
Who searches it?
How long must we retain it?
What evidence value does it have?

■ FINAL

ROVER
Agentic SIEM | SecOps Platform

SIEM IS NOT GOING AWAY

ITS JOB IS CHANGING

From warehouse to lens.

ROVER
Agentic SIEM | SecOps Platform

HOW TO COMPARE PRICING PROPERLY

Do not compare only $/GB. Every platform meters a different job.

- Raw GB or compressed GB?
- Ingested, indexed, analyzed, retained, or scanned?
- Hot retention or cold/archive retention?
- Search included or billed separately?
- Users, detections, SOAR, AI, and data egress included?
- Commitment tiers, overages, and regional rates negotiated?

RIGHT QUESTION

Which architecture gives us the visibility, retention, search speed, and compliance evidence we need at acceptable cost?

SOURCE REFERENCES

Pricing and product details should be checked again before procurement.

- NIST SIEM definition; Cisco/Microsoft/Splunk SIEM explanations
- Vendor pricing pages: Splunk, Sentinel, QRadar, Datadog, Elastic, Sumo, Securonix, Google SecOps
- Security data lake pages: AWS Security Lake, Athena, Sentinel Data Lake, Cribl, Snowflake, Databricks, Panther, Rover
- Pricing changes by region, contract, commitment, feature tier, retention, and negotiated discounts
- Use vendor calculators and quotes before buying decisions

■ Q&A

QUESTIONS & DISCUSSION

ROVER
Agentic SIEM | SecOps Platform

Dhruv Chaudhari
Agentic Cybersecurity | Founder | CTO

What should your SIEM own — and what should your data foundation own?