Security data lake: perché il SIEM si sdoppia tra raccolta e analisi



Il security data lake è la risposta a un problema che i responsabili della sicurezza conoscono bene ma di cui si parla poco: raccogliere e conservare i dati necessari a rilevare un attacco è diventato così costoso da spingere molte organizzazioni a rinunciarvi, cioè a scartare log che sarebbero serviti proprio nel momento peggiore. Il *SIEM*, il sistema che da vent'anni sta al centro del centro operativo di sicurezza, è nato in un'epoca di volumi di dati incomparabilmente più piccoli, e il suo modello economico, che fa pagare in proporzione a quanto si immette, mal si adatta a un mondo in cui i log crescono più in fretta dei budget. Da qui una trasformazione silenziosa ma profonda: la piattaforma non scompare, ma si scompone.

L'idea di fondo è separare due cose che il *SIEM* teneva insieme: la conservazione dei dati e la loro analisi. Da un lato uno strato di raccolta capace di trattenere tutto a costi bassi; dall'altro uno strato analitico che interroga quei dati quando servono. È la stessa logica che ha ridisegnato l'analisi dei dati aziendali dieci anni fa, arrivata ora, con ritardo, alla sicurezza. E non è una moda da fornitori: è la conseguenza aritmetica di un costo per gigabyte che, moltiplicato per i volumi odierni, non regge più.

Perché il SIEM tradizionale non regge più i volumi

Il nodo è il modello di prezzo. La maggior parte dei *SIEM* fattura in base al volume immesso, così ogni fonte aggiunta,

ogni picco di traffico, ogni nuovo sistema da monitorare fa salire il conto. La reazione prevedibile dei team, sotto pressione di budget, è filtrare all'origine: non inviare al *SIEM* i log ritenuti meno critici, campionarli, o conservarli per pochi giorni. Ogni scelta di questo tipo è un risparmio immediato e un punto cieco differito, perché la fonte scartata è spesso quella che, a incidente avvenuto, avrebbe raccontato come l'attaccante è entrato e cosa ha toccato.

È il motivo per cui il problema economico è, in realtà, un problema di sicurezza. Quando l'analista deve scegliere quali dati permettersi invece di quali dati servono, la copertura di rilevamento smette di essere una decisione tecnica e diventa una voce di spesa. La crescita dei volumi, alimentata da *cloud*, identità, *endpoint* e ambienti ibridi, ha reso questa tensione insostenibile, e ha spinto il mercato a cercare un'architettura che disaccoppi il costo della conservazione da quello dell'analisi.

Che cos'è un security data lake

Un *security data lake* è un archivio centralizzato che raccoglie i dati di sicurezza, strutturati e non, su *storage* a oggetti a basso costo, tenendoli disponibili per l'interrogazione senza il sovrapprezzo di un'indicizzazione permanente. La differenza rispetto al *SIEM* classico è che conservare non implica più pagare per analizzare in continuazione: i dati restano lì, economici, e la potenza di calcolo viene applicata quando serve, sui dati che servono. Questo permette di trattenere anni di storico, invece di giorni, e di condurre indagini retrospettive che con la ritenzione ridotta imposta dai costi sarebbero impossibili.

La contropartita è il *tiering*, cioè la distinzione tra dati "caldi", subito interrogabili per il rilevamento in tempo reale, e dati "freddi", conservati a lungo e richiamabili con qualche attesa in più per le indagini e la conformità. Formati

aperti e *storage* di proprietà dell'organizzazione riducono inoltre il rischio di *lock-in*, il vincolo che a lungo ha reso difficile lasciare un fornitore di *SIEM* una volta che ci si erano riversati dentro tutti i log. È un cambio di rapporti di forza: i dati tornano a essere dell'azienda, e la piattaforma di analisi diventa sostituibile.

La security data pipeline: filtrare prima di pagare

Tra le fonti e lo strato di analisi si è inserito un terzo componente, la security data pipeline, che raccoglie i dati, li normalizza, li arricchisce e decide dove instradarli. La sua funzione economica è semplice e potente: ridurre ciò che arriva al *SIEM* costoso, mandando il resto al *data lake* a basso costo, senza perdere nulla. È la promessa che ha reso questo strato l'oggetto delle acquisizioni che hanno riscritto il mercato tra il 2025 e il 2026: CrowdStrike ha rilevato Onum per 290 milioni di dollari, SentinelOne ha rilevato Observo AI per 225, e soprattutto Palo Alto Networks si è presa Chronosphere, una piattaforma di *observability* che porta con sé anche capacità di *telemetry pipeline*: l'operazione, annunciata a novembre 2025 per 3,35 miliardi e perfezionata il 29 gennaio 2026, vale oltre sei volte le due precedenti messe insieme, il segno che rendere sostenibile l'ingestion su larga scala è ormai un asset strategico.

Federated search: cercare dove i dati già sono

La *federated search* è la capacità di interrogare dati che risiedono in posti diversi, un *SIEM*, un *bucket* su *cloud*, una piattaforma di *analytics*, senza doverli prima centralizzare. Serve a superare il presupposto storico che per correlare i dati bisognasse ammassarli tutti nello stesso posto,

presupposto che è anche la radice del costo. Se si può cercare dove i dati già sono, la centralizzazione totale smette di essere obbligatoria, e con essa cade una parte consistente della spesa.

Il mercato si consolida e si biforca

Questa transizione avviene mentre il mercato dei *SIEM* attraversa il suo più grande riassetto. Nel marzo 2024 Cisco ha chiuso l'acquisizione di Splunk per circa 28 miliardi di dollari; a maggio Palo Alto Networks ha annunciato l'acquisto delle attività *SaaS* del QRadar di IBM, perfezionato ad agosto per 1,1 miliardi comprensivi della componente variabile, di cui 500 milioni in contanti, migrando i clienti verso la propria piattaforma Cortex XSIAM; a luglio Exabeam e LogRhythm hanno completato la fusione. E il consolidamento non è solo cronaca finanziaria: il QRadar *SaaS* acquisito da IBM è già stato messo in fine vita, con la prima ondata di dismissioni (QRadar on Cloud, SOAR, Log Insights) scaduta il 14 aprile 2026 e la seconda (EDR, XDR, X-Force Threat Intelligence) fissata al 31 agosto 2026, un prodotto storico uscito dal mercato in meno di due anni dall'acquisizione, mentre le versioni *on premise* restano fuori dal perimetro dell'annuncio.

Sotto il consolidamento si intravede una distinzione più utile per chi deve scegliere, e più sottile della vulgata "aperto contro integrato". Anche gli ecosistemi tutto-in-uno hanno ormai assorbito il *data lake* e la *federated search*, cioè proprio le tecnologie che sembravano la loro alternativa. Microsoft Sentinel, per esempio, ha reso generalmente disponibile un data lake nativo con *storage* a livelli e fino a dodici anni di dati interrogabili, e dall'aprile 2026 può federare dati da Fabric, ADLS e Azure Databricks; la sua piattaforma unisce già *SIEM*, *XDR*, SOAR e gestione dell'esposizione. E non è un caso isolato: anche Splunk, ormai

in Cisco, ha annunciato un proprio *Machine Data Lake* dentro il Cisco Data Fabric presentato a settembre 2025, con una federazione che raggiunge Amazon S3, Iceberg, Delta Lake, Snowflake e Azure, e disponibilità annunciata nel corso del 2026. Lo ha ammesso senza giri di parole Kamal Hathi, a capo di Splunk in Cisco: ingerire tutti i dati dentro Splunk non è un'idea praticabile, la strada è portare l'analisi dove i dati vivono, non costruire un unico grande lago. Il Machine Data Lake non contraddice la frase, la completa: è il livello economico che sta sotto le piattaforme di analisi, da cui i dati vengono promossi solo quando servono. È esattamente la tesi da cui siamo partiti, pronunciata dal fornitore che proprio sul lago unico ha fondato vent'anni di fatturato.

La domanda vera, allora, non è "aperto o integrato", ma due altre: chi possiede lo *storage* in cui vivono i dati, e quanto è portabile il formato in cui sono scritti. Da lì dipende se domani si potrà cambiare piattaforma di analisi senza rifare tutto, ed è la decisione che ogni SOC si troverà a prendere.

Il movimento, poi, non è più solo interno alla sicurezza. Il 24 marzo 2026 Databricks, uno dei nomi del *lakehouse* per l'analisi dei dati, ha presentato Lakewatch, un *SIEM* agentico e aperto oggi in *private preview*, annunciando insieme le acquisizioni di Antimatter e SiftD.ai, quest'ultima fondata dal creatore del *Search Processing Language* di Splunk e dagli architetti del suo *stack* di ricerca. È la conferma letterale del punto di partenza: la logica che ha ridisegnato l'analisi dei dati aziendali non solo è arrivata alla sicurezza, ma ora sono i fornitori del dato, e persino i creatori del linguaggio del vecchio *SIEM*, a entrare nel *SOC*.

Il rischio: un lago senza rilevamento è una palude

C'è un modo di sbagliare questa transizione, ed è trattare lo *storage* economico come un fine invece che come un mezzo. Poter

conservare tutto tenta a fare esattamente questo, cioè accumulare dati con l'idea vaga di analizzarli "poi", e il risultato è un archivio enorme e inerte, sul quale nessuno costruisce rilevamento. Un *data lake* senza *detection* non è una strategia, è una palude: costa meno di un *SIEM* mal dimensionato, ma non protegge di più. Il dato diventa sicurezza solo quando qualcuno scrive le regole di correlazione, definisce cosa cercare e mantiene viva la disciplina del rilevamento, esattamente come nell'evoluzione del SOC verso una difesa che misura la propria efficacia.

Va aggiunta la spinta della conformità, che in Europa rema nella stessa direzione. NIS2 e DORA non fissano una durata minima di conservazione dei log, ma impongono di definirla, documentarla e giustificarla in base al rischio, oltre a garantire l'integrità e la reperibilità delle evidenze, come ricorda la mappa degli adempimenti NIS2: in pratica, la ritenzione lunga smette di essere una scelta discrezionale e diventa una posizione da difendere davanti a un auditor. Su questo un *security data lake* è la risposta economicamente sostenibile, perché consente di trattenere anni di dati senza pagarli come se fossero tutti caldi. Per una volta la conformità coincide con la buona pratica di sicurezza, perché lo storico che serve al revisore è lo stesso che serve all'analista per ricostruire un attacco lento.

Cosa aspettarsi, senza illusioni

Il *SIEM* non sta morendo, si sta scomponendo, e il *security data lake* è la parte di questa scomposizione che ridà alle organizzazioni una leva che avevano perso: decidere cosa conservare in base a ciò che serve alla difesa, non a ciò che il modello di prezzo consente. Il ritorno concreto non è solo il risparmio, ma la fine del compromesso quotidiano tra copertura e budget, e la possibilità di guardare indietro di anni quando un'indagine lo richiede. Resta però la parte difficile, che nessuna architettura regala: il rilevamento, la

correlazione e le persone che li governano. Chi tratta il *security data lake* come una scorciatoia per spendere meno otterrà un magazzino di dati; chi lo tratta come la fondazione su cui costruire un rilevamento finalmente non più razionato otterrà una difesa migliore. La differenza, come sempre, non è nello strumento ma in cosa ci si fa.

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/security-data-lake-siem/