Scaling Security Investigations for Modern Splunk Environments

An architectural guide to modern investigation infrastructure

How security teams scale investigations without replacing Splunk

Featuring BTG Pactual, operating at multi-terabyte scale

btg pactual

imply

Why security investigations don't scale on traditional SIEM infrastructure

Modern security teams rely on Splunk for detection, monitoring, and operational visibility

In mature environments it becomes the backbone of the SOC, powering:

- Dashboards
- Alerts
- Analyst workflows during high-pressure incidents

As telemetry volumes grow and retention requirements expand, **investigation workloads place new demands on SIEM infrastructure.** Security teams must support:

- Large historical lookbacks and
- High-concurrency analysis
- More data without permanently scaling infrastructure

The challenge is not simply data growth

It is the mismatch between steady detection workloads and burst investigation demand

1

Log Volume Growth

Security teams generate more telemetry to reduce blind spots

2

Expanding Retention

Compliance and investigations require longer data retention

3

Bursty Investigations

Incidents trigger high-concurrency searches across historical data

The two primary uses of SIEM

Detection

Continuous monitoring and alerting
Anomalies are investigated

1

Detections

Always-on rules and scheduled searches

2

Events

Signals that something may be wrong

3

Queue

Alerting, triage, and prioritization

4

Analyst

Investigate, correlate, and respond

Investigation

Explain “why” something happened
Explore historical telemetry during incidents

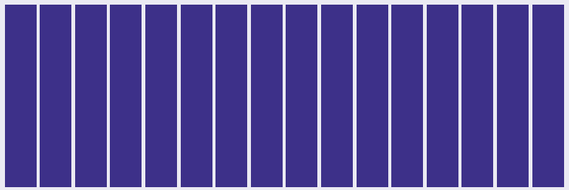
Detection and investigation are fundamentally different workloads

Scheduled searches

Narrow time windows

Always-on compute

Detection Workloads



Time

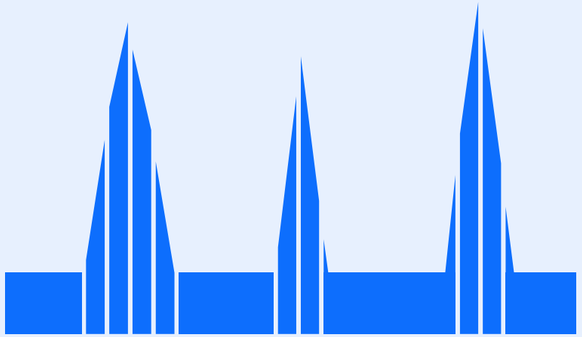
Detection pipelines run continuously to power alerts, dashboards, and automated monitoring. These workloadsare predictable and optimized for speed and reliability.

Ad-hoc searches

Large historical time windows

Bursty compute demand

Investigation Workloads



Time

Investigations behave differently. During incidents, analysts run exploratory searches across large historical datasets often generating bursts of concurrent queries.

Treating these workloads as identical forces organizations to size infrastructure for worst-case investigation demand, even when most activity is steady detection.

Why traditional SIEM optimizations eventually hit a ceiling

1 Reduce Telemetry

- Lower Cost
- Higher Risk

2 Offload to cloud

- Lower Cost
- Slower Access

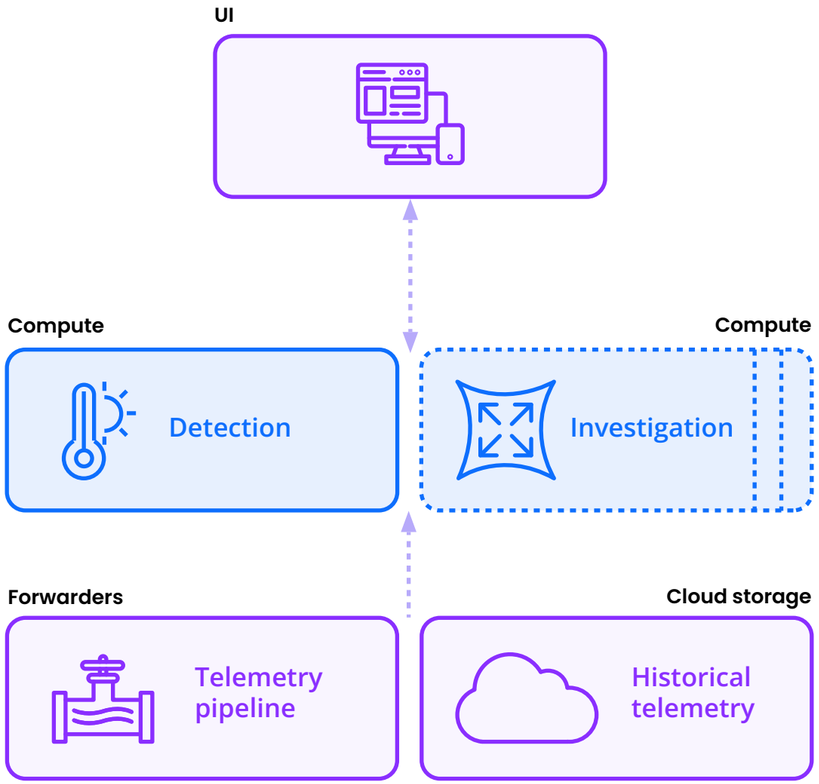
Investigation Blind Spots

80%

Many organizations reduce telemetry to control SIEM costs

Cost savings often come at the expense of investigation visibility

A decoupled architecture for detection and investigation



Detection (Predictable)

Recent, high-frequency telemetry
Continuously indexed for alerts and dashboards
Always-on compute

Investigation (Burst & Elastic)

Full-fidelity historical telemetry stored in object storage
Compute scales only during investigations
No need for continuous indexing

Storage and compute are decoupled. Each workload scales independently.

Case Study

Scaling Investigations at BTG Pactual

When burst investigations outgrew steady-state infrastructure

BTG Pactual relied on Splunk as the backbone of its SOC operations.

Dashboards

Detections

Monitoring

As telemetry grew to multi-terabyte daily ingestion, investigations began to behave differently from monitoring workloads. Incidents required months of historical access, high-concurrency exploratory search, and rapid collaboration across teams.

These were not steady workloads.

They arrived in bursts.

When they did, they stressed infrastructure sized for continuous indexing, not investigative spikes.

The bank was forced into a familiar dilemma

Index everything and overspend

Reduce retention and increase investigative risk

Push data to cheaper storage and accept slow access

Each option traded **cost against visibility.**
None changed how investigations actually behave.

The Realization

This wasn’t a cost problem

It was an architectural mismatch between steady detection and burst investigations

Most organizations consider two paths when costs rise:

Neither approach changed how investigations actually behave.

Optimize Harder

- Reduce indexed volume
- Shorten retention
- Push data to cheaper tiers

Replace Platform

- Reset cost structure
- Rebuild detections and dashboards
- Retrain analysts

- Investigations still arrive in bursts
- Infrastructure still absorbs peak demand
- Tradeoffs between cost, performance, and visibility remain

The problem wasn’t the platform. It was forcing two incompatible workloads to share the same infrastructure.
The answer wasn’t optimization or replacement. It was architectural separation.

Implementing architecture separation with Imply Lumi

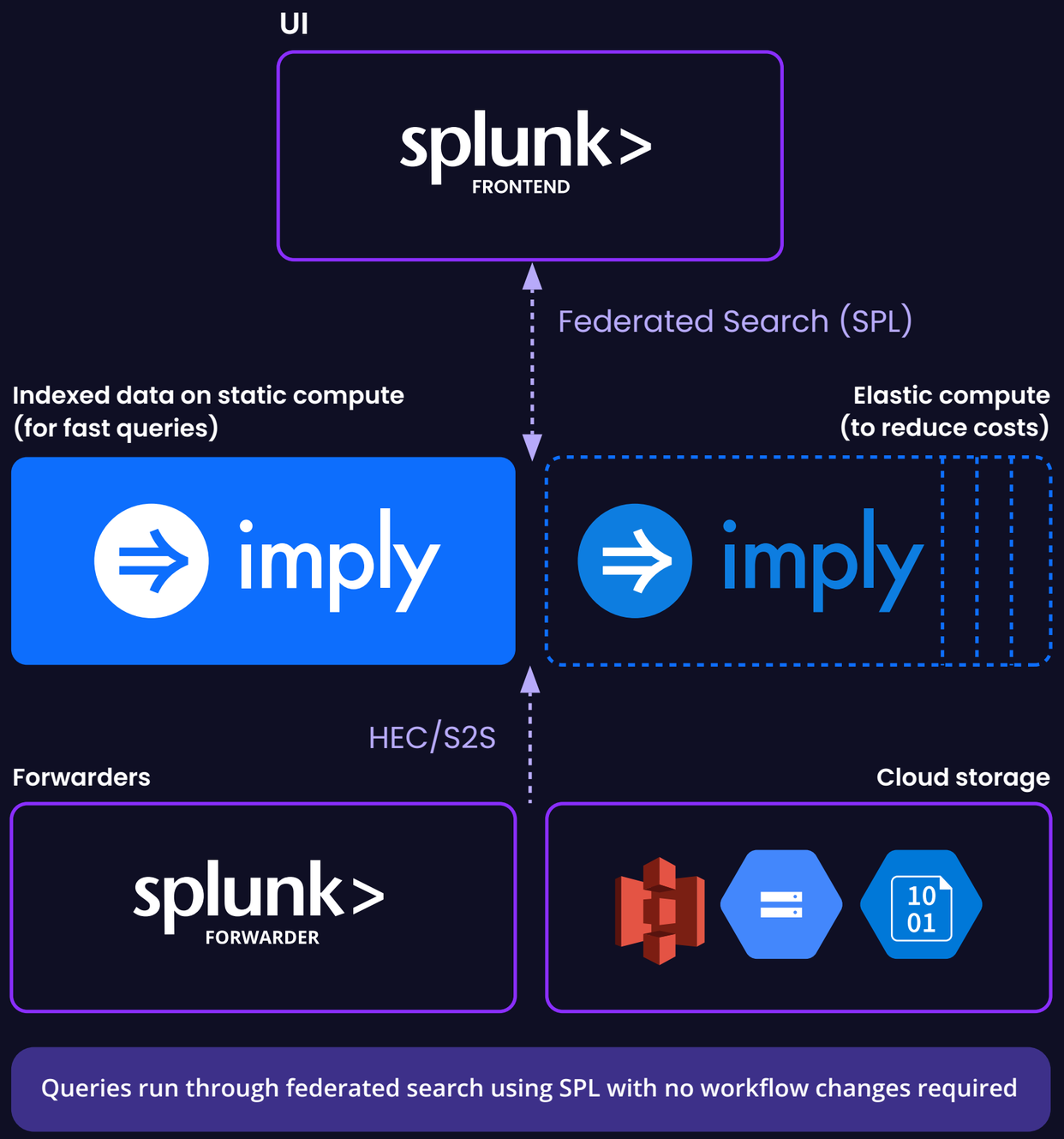
BTG Pactual deployed Imply Lumi as an extension of its Splunk environment.

Detection and monitoring

- Indexed data on dedicated compute
- Optimized for fast, continuous queries
- Powers alerts, dashboards, and workflows

Investigation

Elastic compute scales on demand to analyze large volumes of historical telemetry



What deploying Imply Lumi unlocked

Separating detection and investigation compute changed how the environment scaled.

BTG Pactual replace its SIEM. It changed the architecture underneath it.

+10	4x	60%	Zero
TB/day Data Ingest	Data Retention	Lower Cost per GB	Workflow Changes

Security teams gained scale without sacrificing stability

btg pactual

> We adopted Imply Lumi because we needed to ingest more data, retain it longer, run faster queries, and improve cost efficiency all while keeping our existing dashboards and SPL workflows intact.

Rafael Hass

Security Information Manager
BTG Pactual

Rethinking observability architecture

Investigations require a different approach than steady-state monitoring

Investigations are a different workload

They require elastic compute and deep historical access

Detection and investigation should not share infrastructure

Each workload should run in an environment designed for its behavior

Independent compute unlocks scale

Shared data with separated compute preserves workflows and controls cost

The future of scalable security operations is architectural.

Investigations scale is an architectural decision. Not a tuning exercise.

imply Lumi

An Observability Warehouse

Imply Lumi is an observability warehouse designed to align compute and storage with how workloads behave.

Detection and investigation workloads are decoupled so each can scale independently.

Elastic compute supports burst investigations, while shared storage is optimized for historical search.

Existing Splunk workflows remain unchanged with fine grained control over cost and performance.

Teams keep their tools.
They gain architectural scale.

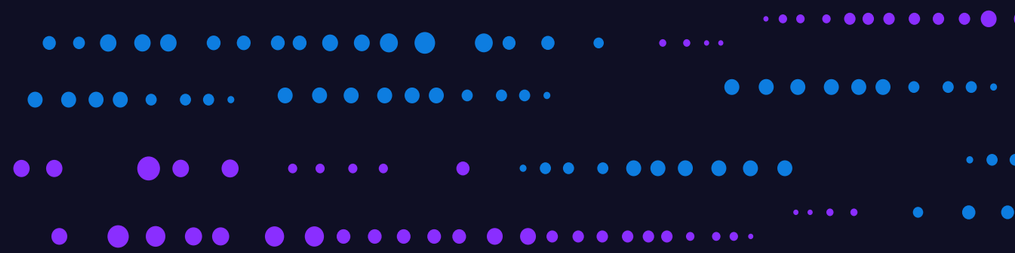
Investigations scale without destabilizing operations.

About Imply

Imply is the company behind Lumi, the modern data layer for observability, security, and AI.

Founded by the creators of Apache Druid®, Imply builds high-performance infrastructure trusted by leadingenterprises worldwide.

Learn more at imply.io.



Most approaches don’t solve the core problem

Optimize	Offload	Replace	Decouple
✓ Reduces indexed volume	✓ Lowers storage cost	✓ Resets cost structure	✓ Compute aligns to workload
✓ Lowers short-term cost	✓ Extends retention	✓ Potential long-term savings	✓ Independent scale
			✓ No workflow changes
✗ Investigations still spike	✗ Slower investigations	✗ Rebuild workflows	
✗ Always-on compute	✗ Limited scale	✗ High operational risk	

The problem isn’t storage or cost.

It’s how infrastructure handles burst workloads.