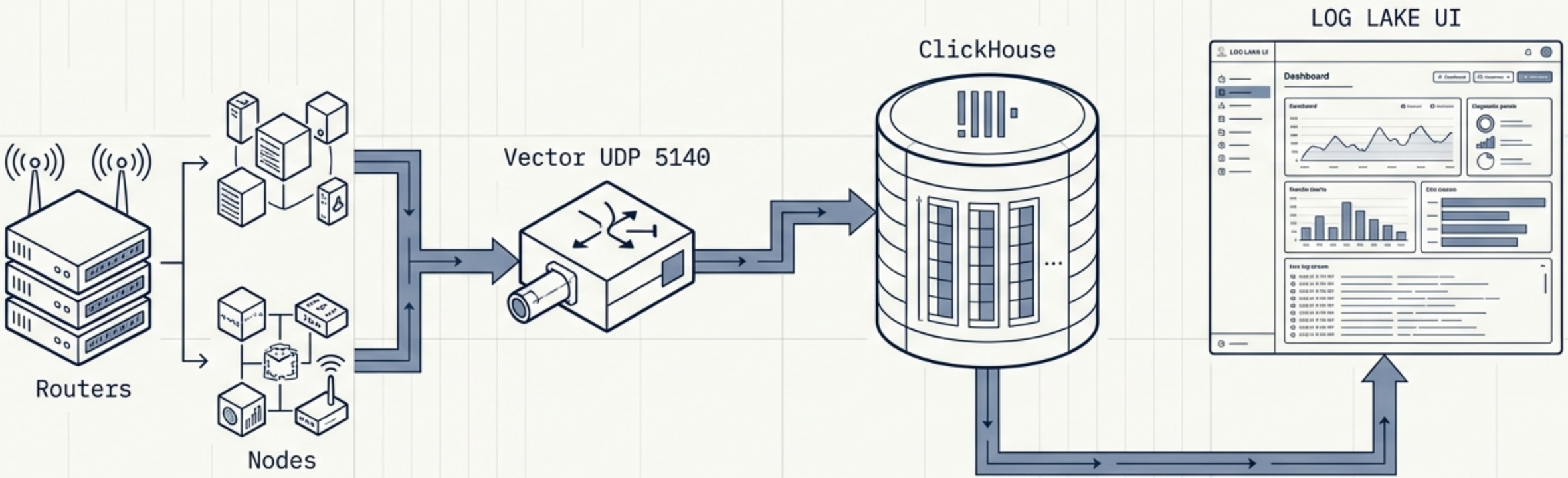
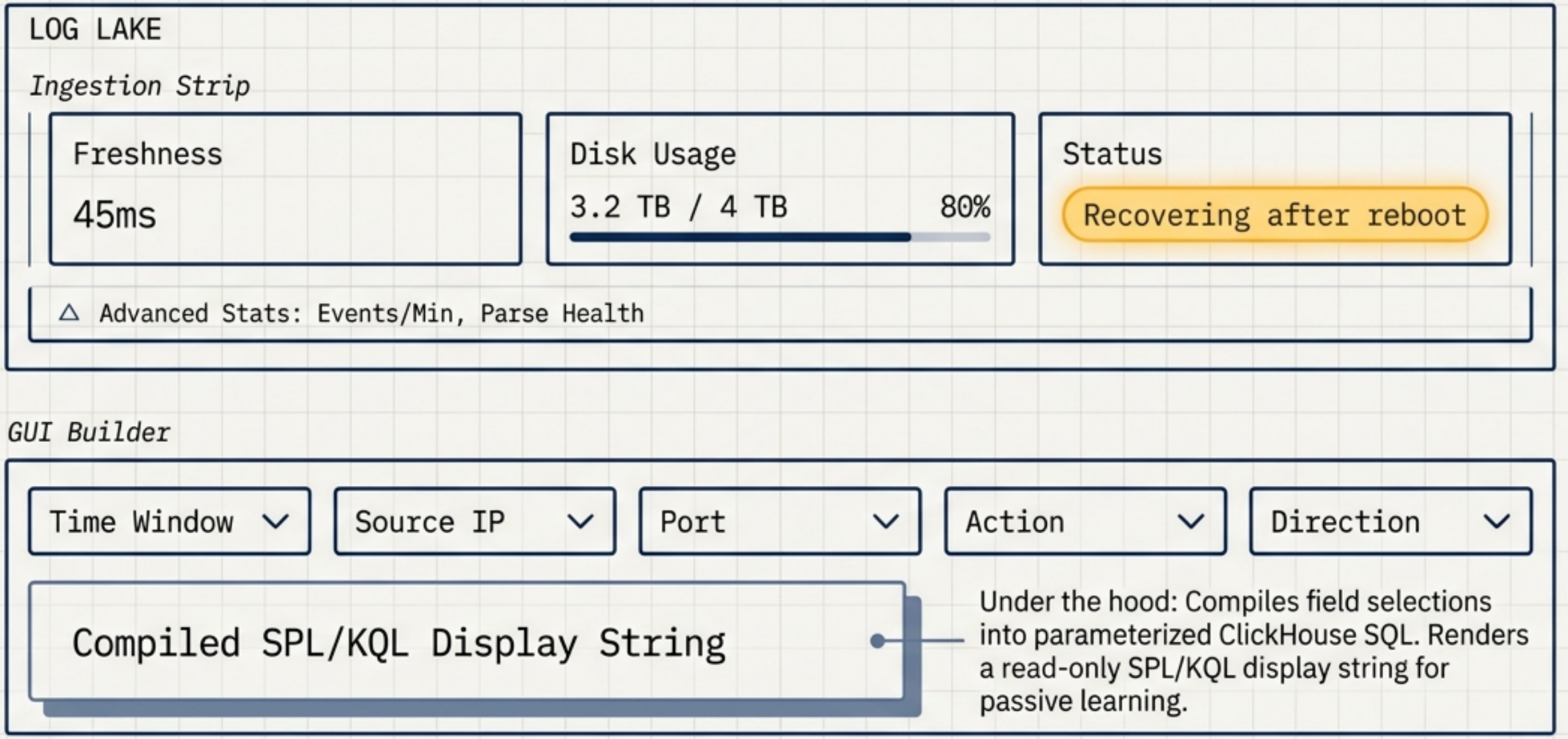
Building LOG LAKE

Architecting and Deploying a GUI-First SIEM for the Home Lab



The GUI-First Homelab SIEM



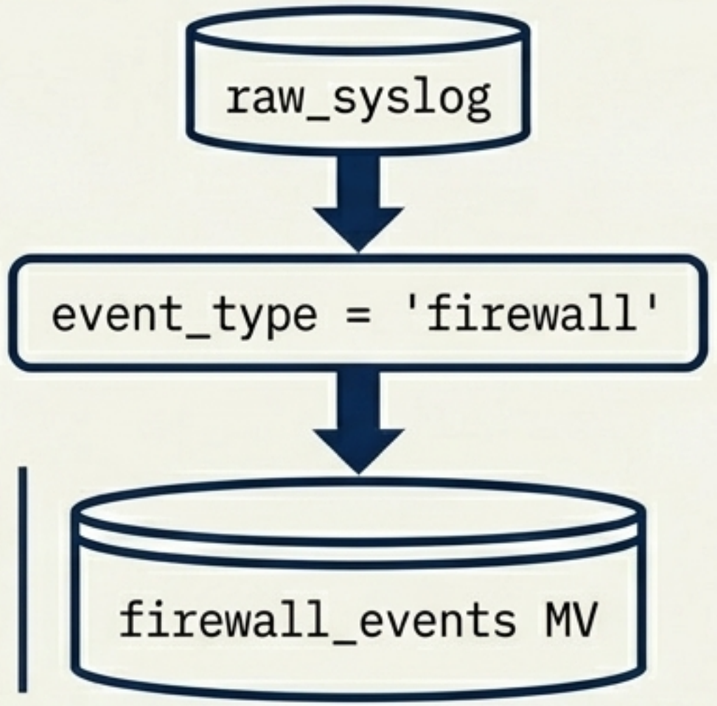
The single operator shouldn't have to memorize custom syntax.
Visual queries generate strict, safe execution.

DB Persona Review: Correcting the `firewall_events` Schema

	Proposed Design	Persona Correction	Architectural Reason
1	`rule_id`	**Removed**	Does not exist on the UDM wire.
2	`action`	**Derived**	Extracted via `multiIf` from the netfilter chain tag prefix.
3	`ports`	`Nullable(UInt16)`	A `CAST` throws on empty maps, rejecting entire insert batches.
4	`ip_addresses`	`Nullable(IPv4)`	Avoids 0.0.0.0 sentinels that corrupt range predicates.
5	`payload`	**Excluded**	Drops `raw/message/fields` credential leak surfaces.

Wave 1: Migration 0016 and the **argMax** Fix

The Pipeline



Ingestion strip caught a VRL bug: missing dpt fields on ICMP-optional ports.

The Code Fix

Yields 37 rows (cartesian product of every reboot).

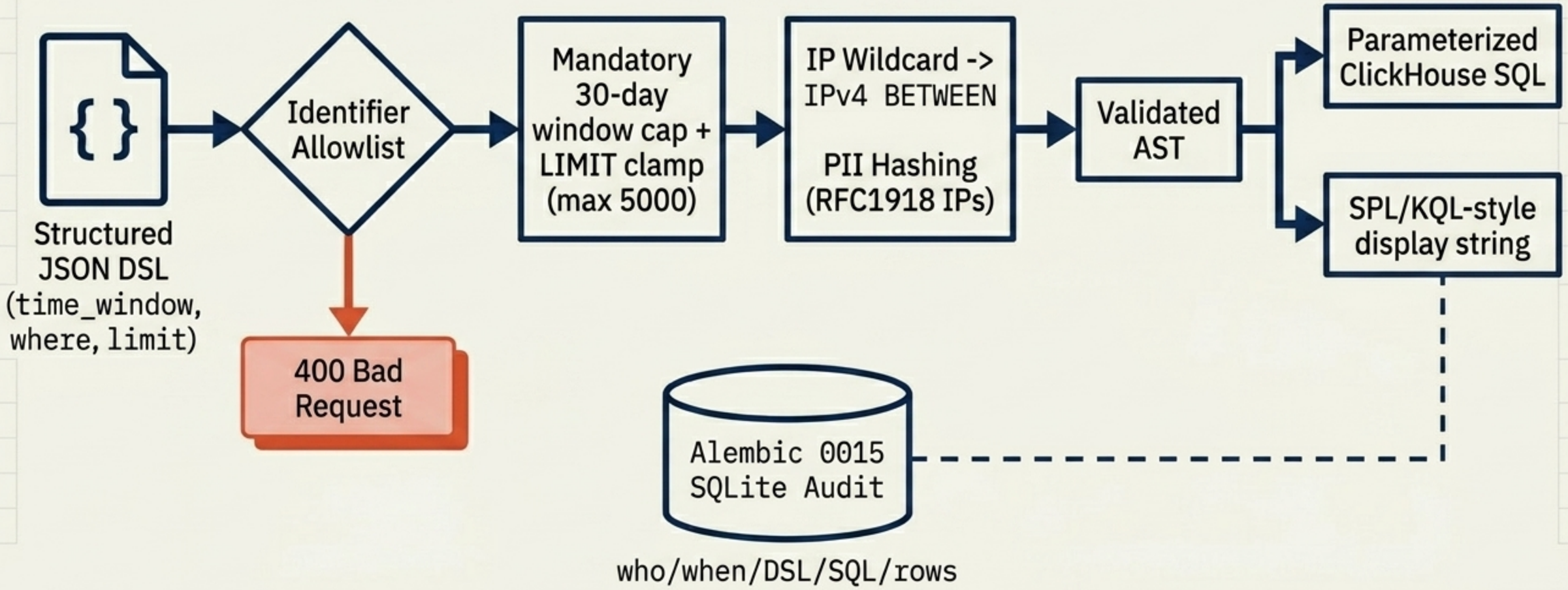
~~**GROUP BY** (source, host, component)~~

argMax(host, ts) and **max**(ts) **AS** last_ts

Yields exactly 1 row per source.

The nested-aggregate trap: Aliasing **max(ts) AS ts** triggers a ClickHouse parser error because **ts** is already an aggregation key. Always rename to **last_ts**.

The Safety Harness: Single AST, Dual Output



The Deploy Afternoon Autopsy

```
[SUCCESS] 1193 backend pytest passed.
[SUCCESS] 351 vitest passed.
[SUCCESS] Playwright e2e full pass.
> Executing deploy to live Mac mini...
[WARNING] 5 production-only bugs surfaced.
```

At 3:41 PM, the code shipped. Then, **five bugs surfaced**—one after another—none of which CI could ever catch.

Bug 1: Two Layers Fighting for Readonly

[Layer: ClickHouse]

Symptom: Immediate HTTP 500 on **POST /api/siem/query**.

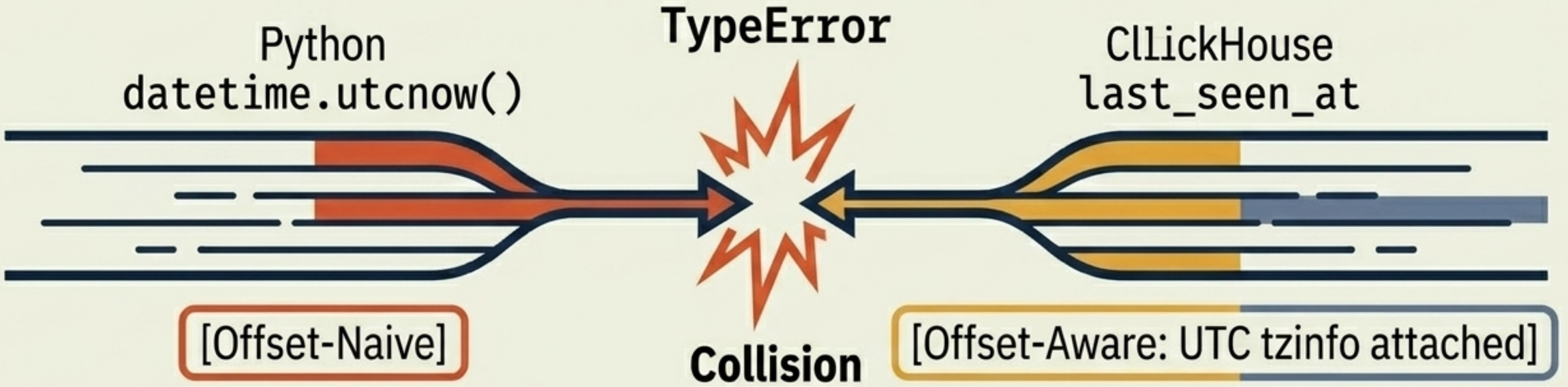
```
ProgrammingError from
ClickHouse driver.
Traceback (most clack list):
  File ".", line 4, in <module>
  File "/toot", line 37, in pumdlts
    starter_ look()
  File "/readony", line 153, in not
    return al(renornany)
```

The server enforced a **readonly** profile via `clickhouse-users.xml`.

The python client independently passed `settings={readonly:1}`.
A server-readonly session outright rejects client-set settings, triggering a crash.

Drop the connect-time settings dict.
Let the server enforce it natively.

Bug 2: The Naive vs. Aware Disconnect



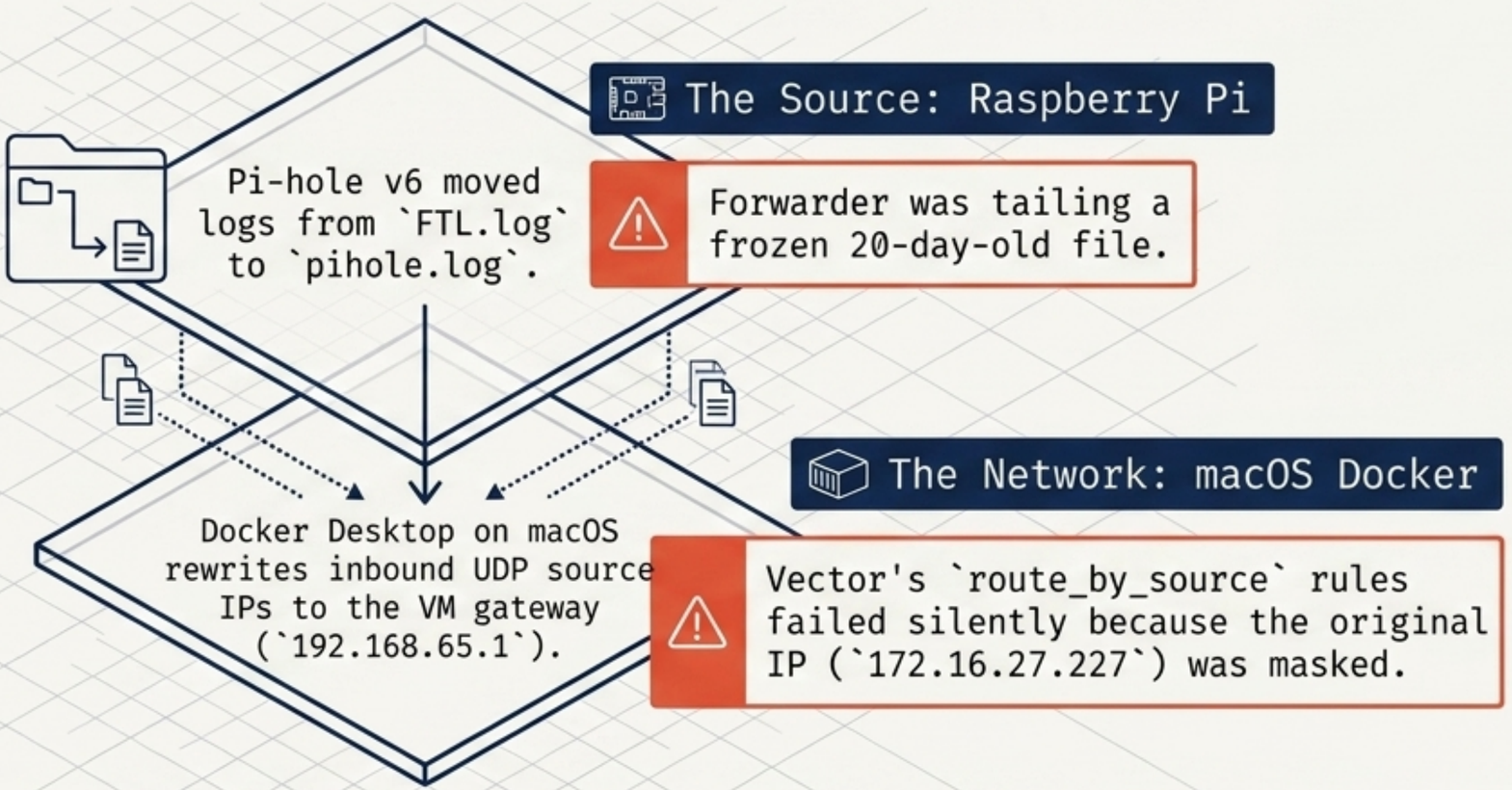
Symptom

`poll_security_finding_emit` crashing every 5 minutes. A silent failure pre-dating LOG LAKE.

Fix

Implement a `naive_utc_now()` helper. Normalize both sides to naive UTC before doing subtraction arithmetic.

Bug 3: Stacked Failures on Pi-hole Syslog



Fix:

Update the file path on the Pi (Layer A). On Vector, route by the hostname in the syslog payload instead of the IP in the UDP frame (Layer B).

Bug 4: The Inode Cache Trap

Host View		Container View
File: /etc/vector/vector.yaml	≠	File: /etc/vector/vector.yaml
Result: 547 lines		Result: 535 lines

Symptom Vector container healthcheck failing for 28 hours.

Root Cause macOS Docker Desktop bind-mount caching. The file was updated on the host, but the container read a truncated, stale copy.

The Trap Standard inode rewriting (writing to tmp, then mv over target) failed to bust the macOS cache.

The Fix

```
docker compose up -d --force-recreate vector
```

Bug 5: Three Bugs Hiding in One Timestamp

Bug 5a: Timestamp Skew

UDM sends zoneless BSD timestamps. Vector's global syslog timezone assumes UTC local time, skewing data 4 hours into the future.

⚠ Fix: Trust Vector's .time_received.

Bug 5b: Doubled Hostname

UDM sends 'UDM-Pro' twice in the frame. Vector consumes the second as the appname, breaking the sequence and losing the [TAG].

⚠ Fix: Make the [TAG] optional.

Bug 5c: The Unknowns

Even after fixing parsing, 159,909 rows logged with action = 'unknown'.

Dodging the Bytes-Codec Trap

159,909 rows of action = 'unknown'

Path A - The Vetoed Rewrite

Build complex bytes-codec syslog source

Parse doubled frames

Recover netfilter tag

Expensive, High Risk

The Revelation

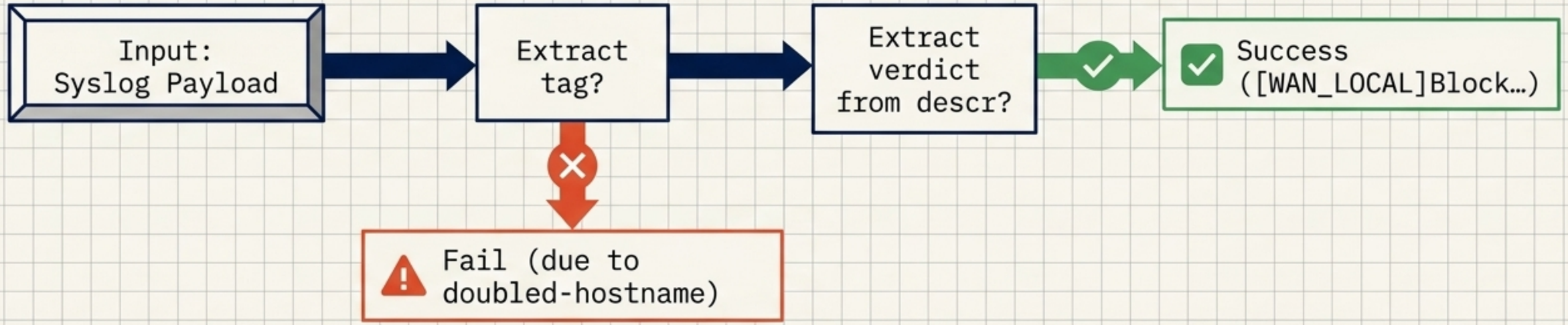
100% of rows were inferred = 'dnat'. The UDM wasn't logging Drops/Accepts by default on the WAN! Path A would have perfectly parsed useless data is perfectly parsed useless data.

Path B - The Redirect

Run live cardinality query on firewall_events

The Real Fix: Check a single box in the UniFi UI for per-rule logging.

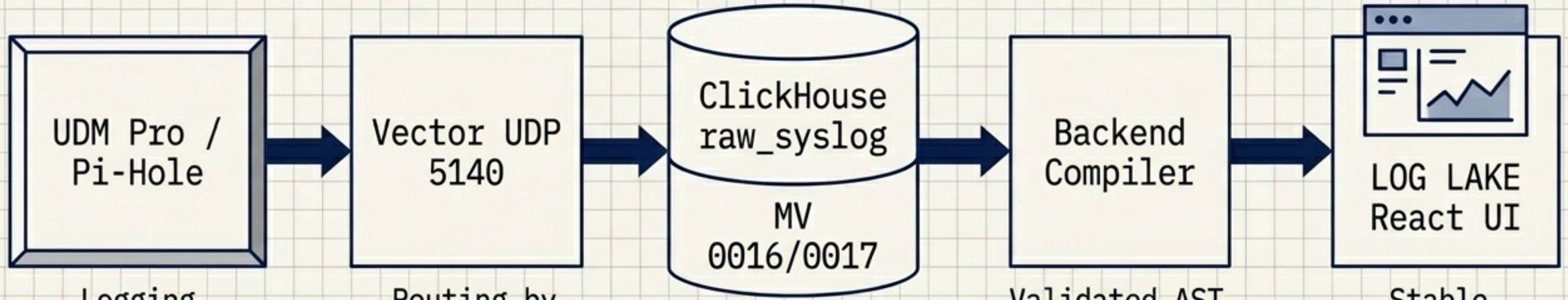
Migration 0017: The descr Fallback



Metric	Before 0017	After 0017
Distribution (Last 10m)	221 rows, 100% unknown	~50% dnat, ~50% drop, 0% unknown

Architectural Philosophy: Forward-only. Do not run heavy IO to backfill 180 days of raw_syslog. **Let the 159,909 old rows age out via TTL.**

LOG LAKE: Post-Deployment Architecture



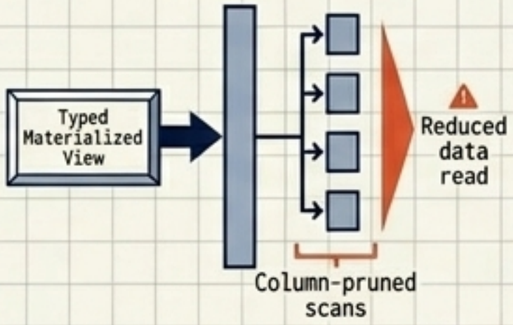
5 deployed bugs resolved. 4 pre-existing bugs eliminated. Pipeline stabilized.

Architectural Lessons Learned

Typed MVs > Parameterized Queries

Migration effort pays off immediately via column-pruned scans.

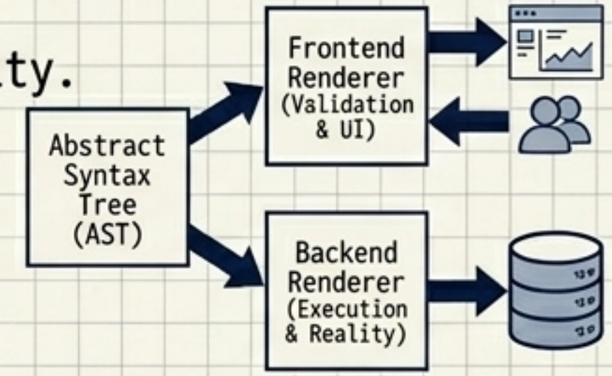
Structure the data once, read it fast forever.



One AST, Two Renderers

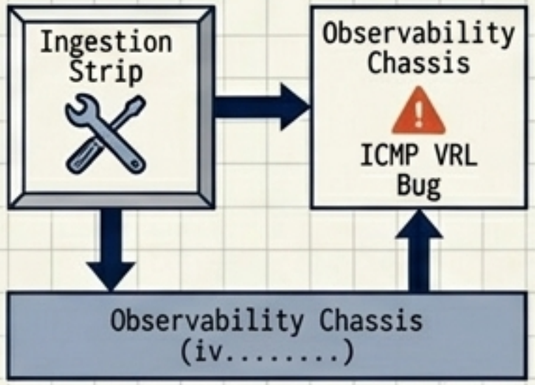
Frontend validation mirrors backend reality.

Users read what they don't have to write, learning syntax passively.



Test Rigs in Production

The ingestion strip caught its own bugs (ICMP VRL).
Build observability directly into the chassis of the application.



Query Before Surgery

Always run live cardinality before executing complex parser rewrites.

Define the reality of the data before trying to fix it.