ClickHouse Log Aggregation

ClickHouse Log Aggregation & Investigation

Architecture Overview

Log Pipeline (Migrated to Vector - Feb 2026):

- **Terra (server):** Journald + Docker containers → Vector
- **Luna (backup server):** Journald → Vector
- **Pluto (gateway):** Journald + nftables (firewall) → Vector
- **Vector Pipeline:** Unified Jinja2 Ansible template. Converts all logs to **America/New_York (EST)** before storage.
- **All sources** → ClickHouse (system_logs table with embeddings)

ClickHouse Access

- **URL:** https://logs.flora.family
- **User:** flobot
- **Database:** default
- **Table:** system_logs

Table Schema

```
CREATE TABLE default.system_logs (
  timestamp DateTime64(3),
  host String,
```

```
  unit String,
  message String,
  priority Int8,
  log_id String,
  embedding Array(Float32),  -- 768-dim vectors
  source String DEFAULT 'systemd'
)
ENGINE = ReplacingMergeTree
ORDER BY log_id
```

Log Sources

Vector Migration (Feb 2026)

- **Retired Fluent-Bit:** Entire lab moved to Vector for better performance and easier Ansible management.
- **EST Alignment:** All timestamps across the pipeline are now surgically aligned to Eastern Time.
- **Container Naming:** Docker logs now show friendly names (e.g., `ollama`, `bazarr`) instead of raw hex IDs.

Noise Management & Filtering

- **Ollama/GIN:** Sampled 1:100 to prevent database wear.
- **Sanoid:** Running in `--quiet` mode on Terra and Luna.
- **Recursive Loops:** Patched VRL filters to prevent `log-brain` from logging its own status updates to ClickHouse.
- **Context Truncation:** `embeddings_worker.py` now truncates logs at 3,072 characters to stay within Ollama context windows.

Embedding Pipeline

Performance (Feb 2026):

- **Coverage:** Reached **100% vectorization coverage** for all essential services.
- **Model:** nomic-embed-text (768-dimensional)
- **GPU:** Quadro P2200 for processing.

Key Learnings

1. **Timezone consistency is critical:** Migrating from UTC to local time in the pipeline saves massive mental overhead during investigations.
2. **Priority Intelligence:** Mapping systemd priorities (0-7) into the database allows for instant "Error-only" dashboards.
3. **Recursive logging is a DDoS:** Always filter your log-aggregator's own logs out of the stream.
4. **ReplacingMergeTree quirk:** Must insert complete rows (all columns) for deduplication to work.

Future Work

- ☑ Add luna backup server logs
- ☑ Align all hosts to EST
- ☐ Parse firewall log CSV columns for structured queries (Pluto implementation)
- ☐ Real-time anomaly detection via vector search
- ☐ Log retention policies & archival

Revision #3
Created 2026-02-17 02:44:56 UTC by Flobot
Updated 2026-02-27 02:22:23 UTC by Flobot