Medicaid Fraud Hunter

Investigative Pipeline — Executive Summary

May 2026 | Sean Yunt

Purpose

Medicaid Fraud Hunter is a self-hosted analytical pipeline that scans publicly available HHS Medicaid claims data to identify providers exhibiting statistically anomalous billing patterns. It produces ranked suspect lists and evidence-grade PDF dossiers suitable for attorney review and regulatory referral.

The tool operates entirely on-premises, requires no cloud services, and processes the full national Medicaid dataset — over 617,000 providers and 159 million procedure rows — on commodity server hardware.

Infrastructure

The pipeline runs on a single Dell Precision T3600 workstation located on-premises:

Component	Specification
Server	Dell Precision T3600 Workstation
CPU	Intel Xeon E5-1650 — 6 cores / 12 threads @ 3.2GHz
RAM	64GB ECC DDR3 (required for national dataset processing)
GPU	NVIDIA GeForce GTX 1060 6GB (CUDA acceleration)
Storage	120GB NVMe SSD (OS) + 1TB HDD (data)
OS	Debian Linux
Runtime	Docker containers
Total Hardware Cost	$237

All pipeline components run as Docker containers with automatic restart on reboot. PDF dossiers are served via an on-premises Nginx file server accessible at http://10.0.0.45:8083/dossiers/.

Data Sources

Dataset	Source	Size	Contents
Medicaid Provider Spending	HHS / data.cms.gov	2.8 GB	All Medicaid claims — 617,503 providers, 159M procedure rows
NPPES Provider Registry	CMS National Plan & Provider	1 GB (zip)	Provider names, addresses, specialties, NPI numbers

Both datasets are publicly available and updated periodically by CMS. The pipeline must be re-run when new data is published.

Pipeline Stages

The workflow consists of three sequential stages. Stage 1 is run once per dataset version. Stages 2 and 3 are run on demand.

Stage 1 — Preprocess (One-Time)

Reads the raw 2.8GB claims file and aggregates it into two optimized summary tables:

- provider_monthly.parquet — 202 MB, 20,074,086 rows — monthly billing totals per provider
- provider_procedure.parquet — 793 MB, 159,437,730 rows — procedure-level detail per provider

These files eliminate repeated full-dataset reads on every subsequent scan. Without preprocessing, each scan would re-aggregate 159 million rows from scratch.

Metric	Value
Input	2.8 GB raw parquet file
Output	~1 GB in two summary parquet files
Runtime	~6 minutes
Frequency	Once per dataset version
Peak RAM	~35 GB

Stage 2 — Scan (On Demand)

Runs four independent anomaly detectors against all 250,825 qualifying providers (those billing $100,000+ total). Results are ranked by corroborating evidence score and saved to a CSV file.

Anomaly Detectors:

Detector	What It Catches	Method	National Flags
Volume Impossibility	More than 1,500 claims in a single month	Hard threshold	59,257 providers
Revenue Outlier	Abnormally high revenue per claim vs. peers	Median/MAD comparison	27,675 providers
Billing Spike	Sudden 5x+ surge vs. provider's own history	Rolling baseline comparison	8,478 providers
Suspicious Consistency	>90% of bills at identical dollar amount	Consistency ratio	0 providers*

* Suspicious consistency flagged 0 providers at national scale — threshold calibration pending.

Scoring logic — providers are scored based on the number of independent detectors that fire:

- 1 detector: maximum score of 70%
- 2 detectors: maximum score of 90%
- 3+ detectors: score of 100%

Metric	Value
Providers analyzed	250,825 (366,678 excluded below $100K threshold)
Providers flagged above 0.3 threshold	79,944
Top score	85% (3 corroborating detectors)
Runtime	~3.5 minutes (using preprocessed data)
Output	scan_results.csv — ranked list with NPI, score, flags
Frequency	Once per dataset version, or per targeted state query

Scans can be filtered by state for targeted investigations:

```
scan.sh --state WV
```

Stage 3 — Profile (On Demand, Per Provider)

Generates a detailed PDF investigation dossier for a specific provider NPI. This is the primary deliverable for attorney review.

Each dossier contains:

- Provider identity — name, address, specialty, NPI (from NPPES registry)
- Claims summary — total claims, total paid, beneficiary count, date range

- Top procedures by volume and dollar amount
- Peer comparison — percentile rank and z-score against all 617,503 providers nationally
- Monthly billing timeline — month-by-month claims and payment totals
- Detected anomalies with severity and supporting evidence

Metric	Value
Runtime	Under 2 minutes per provider
Output	PDF dossier at /mnt/storage/dossiers/
Access	http://10.0.0.45:8083/dossiers/ (local network)
Frequency	On demand — as many times as needed

Provider Lookup

Before profiling, a provider can be identified by name using the NPPES registry lookup:

```
lookup.sh "Acme Health Clinic" --state WV
```

Returns the provider's NPI, address, and specialty — which is then passed directly to profile.sh.

End-to-End Investigative Workflow

Typical workflow when a specific provider is reported as suspicious:

Step	Command	Output
1. Identify provider NPI	lookup.sh "Clinic Name" --state WV	NPI number + address
2. Generate dossier	profile.sh <NPI>	PDF dossier
3. Review dossier	http://10.0.0.45:8083/dossiers/	Open in browser
4. Refer to attorney	Share PDF	Qui tam referral

Typical workflow for proactive investigation from scan results:

Step	Command	Output
1. Run national scan	scan.sh	scan_results.csv — 79,944 flagged providers

Step	Command	Output
2. Review top suspects	Open CSV	Ranked list with scores and detector flags
3. Profile top candidates	profile.sh <NPI>	PDF dossier per provider
4. Refer to attorney	Share PDF	Qui tam referral

Key Numbers at a Glance

Metric	Value
Total providers in dataset	617,503
Providers analyzed (above $100K threshold)	250,825
Providers flagged nationally	79,944
Top suspect score	85% (3 corroborating detectors)
Preprocess runtime	~6 minutes (one-time)
Scan runtime	~3.5 minutes
Profile runtime	Under 2 minutes
Total infrastructure cost	$237
Estimated AWS equivalent cost per scan	~$0.76 on-demand / ~$0.23 spot
On-premises cost per scan (electricity)	~$0.04

Current Limitations

- Suspect identification is algorithmic — findings require human verification and legal review before any referral action
- Suspicious consistency detector requires threshold recalibration for national scale (currently flags 0 providers)
- Provider dossier risk score display bug — shows 0% instead of scan-derived score (fix pending)
- No automated alerting when new HHS datasets are published — manual re-run required
- Qui tam referral requires partnership with a False Claims Act attorney for legal execution

Pending Enhancements

- Fix dossier risk score to reflect scan-derived score
- Recalibrate suspicious consistency detector for national scale
- Add automated dataset update detection
- Evaluate PySpark scan for parallel multi-state processing

This document is confidential and intended for internal use only.