NETSCOUT

From Chaos to Efficiency: Driving AI Scalability Through Data Normalization and Token Reduction

Optimizing AI with Structured Data and Token Management

May 2026

Richard Fulwiler

Senior Director Product Management, Service Providers

“We’re asking AI to explain network behavior — but we’re feeding it data that was never designed to explain anything”

Most of the data we feed AI was built for Human Dashboards, Threshold Alerts, Post-incident Reporting

AI Sees fragmentation instead of sessions, symptoms instead of causes, what happened instead of why

AI Sounds Confident, acts fast – and still gets the answer wrong

Why Alarms, Logs & NEM's Events Mislead AI

They Report Symptoms, Not Causes – and Teach AI the Wrong Lessons

⚠ What Alarms & NEM's Events Give AI

Threshold-Driven, After-the-Fact

- Fire only after user impact — pre-failure degradation is invisible; alerts suppress under load exactly when AI needs signal most

Alarm Storms Obscure Root Cause

- One failure generates hundreds of correlated alarms; AI learns correlation, not causality — symptoms get ranked, not explained

No Subscriber or Procedure Context

- Reports "Attach Failure" without which procedure, subscriber, or retry failed — AI sees WHAT crossed a threshold, never WHY

Vendor-Filtered and Siloed

- Each NEM exposes only what its vendor chooses — cross-domain failures disappear; AI reasons over partial truth, confidently wrong

✅ What Packets Reveal

On-the-Wire Ground Truth for AI

- Captures every packet on every 3GPP interface + RAN traceport — vendor-agnostic; what the network actually did

Full Causal Chain

- Shows which protocol step failed, on which interface, in what sequence — NAS, SIP, Diameter, PFCP, GTP-U correlated across RAN, Core, and User Plane

Subscriber-Keyed Evidence

- Every failure tied to a specific subscriber, device, slice, and cell — distinguishes network fault vs. policy, interop, RAN, or handset issues

See Failure Before Alarms Fire

- Latency spikes, retransmissions, malformed messages — root cause is visible before thresholds are crossed

⚠ **Alarm only sees:** "Attach failure" ➜ ✅ **Packets reveal:** NAS malformed at N1 → AMF rejected auth → UE retried 3x → session abandoned — Cause, sequence, and fix path explicit

The Scale Challenge: Packets Are Massive

A 5G SA Network with 10M Subs Generates ~ 9 PB/day of Data

~9.035 PB/day

Total Data Produced Per Day

0.9 GB per subscriber per day

User Plane — Probe Ingestion

~9 PB / day

Probe Ingestion per day

Control Plane — NAS / SBI Signaling

~10 TB / day

Every Packet on every 3GPP interface

RAN — Traceport Records

~25 TB / day

Per-UE RRC, radio measurements, beam data

You cannot feed raw packets into AI — this is exactly what curation solves

Raw Packet Data Breaks AI — Curation is the Fix

Curation Transforms Fragmented Packet Data into AI-Ready Causal Intelligence

Granularity Breaks AI Reasoning

Billions of fragmented packets overwhelm context windows — 60–80% of AIOps compute wasted on data prep

Identity Fragmentation Hallucinates

Temporary IDs (GUTI, TEID) shift mid-session — AI misattributes failures and fabricates root causes

Cross-Domain Causality Lost

RAN, Core, User Plane events are siloed — AI mistakes correlation for causation, wrong team gets blamed

Curation Engine

Procedure-Level Records

Packets converted to semantic records — causal meaning preserved at a fraction of the volume

Reconstruct & Correlate

Transaction reconstruction reassembles control-plane sequences and binds user-plane flows. Cross-domain correlation (CP+UP+RAN) preserves temporal ordering and causal dependencies

Extract & Normalize

Derives protocol/state attributes, KPI metrics, and procedure markers. Multi-protocol normalization maps results into schemas across NAS,SIP, Diameter, PFCP, and GTP-U

Stable Subscriber Identity

Resolves shifting IDs into stable SUPI/IMSI — precise cross-domain (RAN& CORE) correlation without hallucinated ownership

Volume reduction solution sacrifices reasoning: collapsing sessions strips protocol sequence, leaves subscriber IDs unresolved, and removes the cross-domain context causal AI requires

Curation + Contextual Intelligence: The Secret Sauce

Curation structures facts. Context Intelligence assigns meaning

INPUT

3GPP Packets (From Packet Core Virtual Sensors + RAN traceport)

Fragmented, anonymous, disordered

STEP 1

CURATION

Structural transformation — makes data coherent

- Normalizes packets into subscriber-keyed procedure records
- Resolves shifting IDs (GUTI, TEID) into stable SUPI/IMSI
- Extracts Location, Timing, Session, Domain and KPI dimensions

Produces: Structured Network Facts

STEP 2

CONTEXTUAL INTELLIGENCE

Semantic enrichment — gives facts meaning and cause

- Access Context: stable identity, NAS procedures, RAN metrics
- Core Context: policy decisions, cause codes, slice/domain KPIs
- User Plane Context: per-subscriber QoE, app, cell, and slice attribution

OUTPUT

AI-Ready Causal Intelligence

↓ *Model Context Protocol Server*

↓ *AI Agents*

Together, AI stops guessing

What You Get: Intelligence + Economics

Curation doesn't just reduce volume — it delivers structured intelligence and drastically lower AI inference costs

Telecom Ontology — Built In

Every fact organized into typed entities and causal relationships — UE, session, slice, cell. AI understands the network as a domain, not just fields

Service Graph — No Model Required

Session paths are read directly from the 3GPP signaling at query time. Nothing to build or maintain. Any static graphs is stale the moment a handover occurs

Blast Radius Analysis — Instant

Impact scope correlated analytically across subscribers, cells, and slices at query time — observed data, not modeled relationships. Faster and never stale

Third-Party AI — No Retraining

Curated data consumable by any AI — LLMs, agents, third-party tools — without custom preprocessing or domain-specific retraining

Economics Delivered

>90%

Observed Token Reduction

Real Savings $

Could be a Larger effect for Sovereign and Domain Models

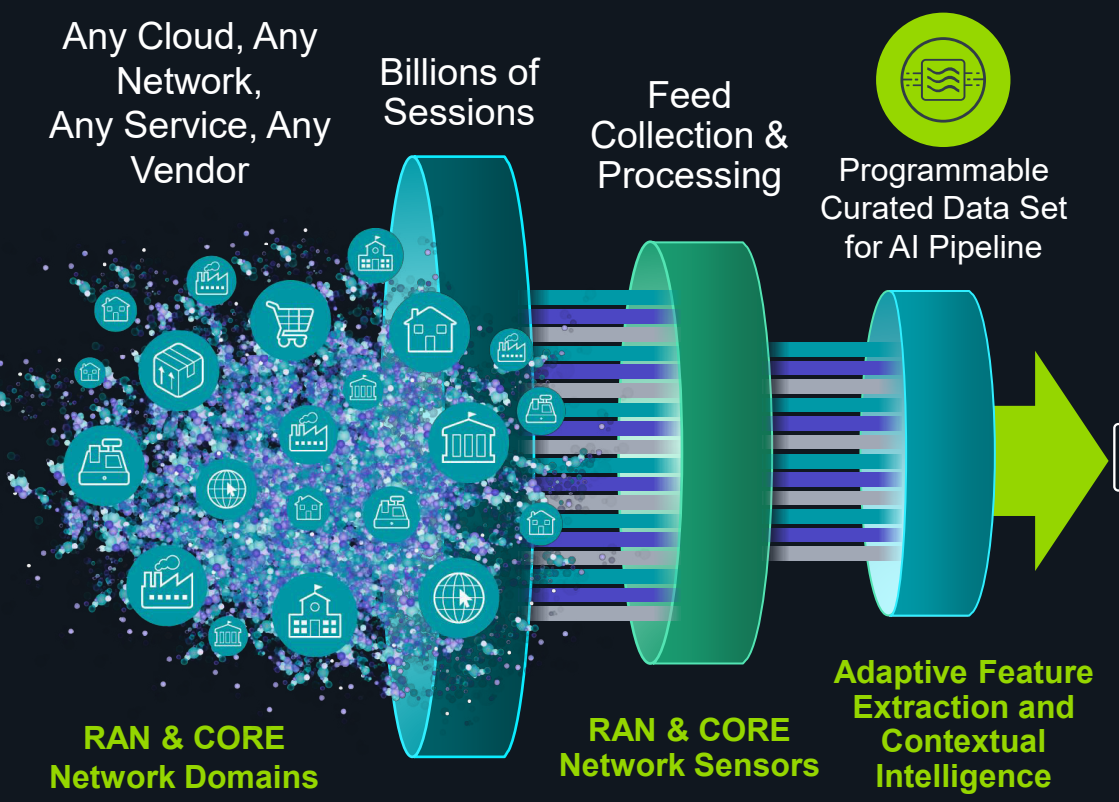
Direct impact:

- Fewer GPUs / Lower memory footprint
- Higher agent concurrency at the same price
- Unlocks continuous AIOps — not just batch processing

You don't build these separately — they emerge from curation itself

The Pipeline: Sensors to AI Agents

Collect → Curate → Enrich with Context → Deliver Smart Data to AI



Curated Data for AI-ready Intelligence

Preserves the meaning through ontology-driven feature extraction, aggregation, and labeling of metrics, entities, and relationships—feeding the AIOps pipeline with Smart Data from RAN, Core, MEC, and Transport

Telecom Experience Feed AIOps

Enhanced Subscriber Insights packaged with Telecom experience to pinpoint complex RAN/CORE/MEC/Transport issues

AI Driven Use Cases

Monetization Use Cases, improve NPS, Network Exposure API's, Prediction, Planning Slice SLA management, Digital Twin generation, Mobile Threats

Generate once, feed everything

Curation in Action: Two Use Cases

Same curated pipeline — no custom data prep, no raw packets, no alarm storms

Mobile Banking

Session integrity for real-time transfers

CSP banking app needed to verify subscriber location and session integrity for international transfers. Direct HLR/VLR queries couldn't handle real-time transactions. Packet streams overwhelmed the app with too much raw data

MCP curation layer: Clean, low-latency interface delivers only what the application needs — subscriber location, session state, and integrity in a single query

0ms → Query · 8ms → MCP lookup · 45ms → Facts returned

Transfer approved in <100ms

Agentic AI Network Triage

VoNR call setup failures, Dallas cell cluster

t=0s Failures exceed threshold → curation engine publishes procedure-level facts + subscriber list to MCP via topic/channel

t+2s Agent auto-triggered — queries MCP for RAN conditions, procedure failures, KPI trends

t+8s Failures isolated to single AMF instance, correlated to recent policy change — fault ownership assigned automatically

t+12s Root cause identified, right team notified with fix path explicit. No human escalation chain

12 seconds Vs **45+ min** human escalation

Both enabled by the same curated pipeline — generate once, feed everything

The Power of Seeing Differently

Smart Data Enables AI That Actually Works

Start with Ground Truth Data — 3GPP Packet Core & RAN Traceport

Curation is not Optional to Drive Token Economics

Structure + Contextual Intelligence is Reasoning Fuel for AI

MCP delivers Curated Facts to AI Agents

Flow Aggregation Strips the Story, You See The Crash, Never the Chain of Events that Caused it

Intelligence is not about how much data you collect — it's about how well you make AI understand it

NETSCOUT®

Guardians of the Connected World