Google Cloud

Autonomous Network Operations

Transforming Telecom with AI and Google Cloud

Authors : Yogesh Tewari, Naresh Rao

Contributors : Greg Cheng, Shishir Kurhade, Manish Gupta

Introduction

Traditional network operations are often manual, reactive, and struggle to keep pace with the dynamic nature of modern telecom networks. Autonomous Network Operations aims to address these challenges by introducing a high degree of automation, driven by AI and machine learning. This allows for proactive network management, predictive maintenance, and optimized resource allocation. Google Cloud provides a rich ecosystem of tools and services that are ideally suited to build and deploy such a sophisticated Autonomous Network Operations framework.

Executive Summary

The Challenge

Communication Service Providers (CSPs) face a critical inflection point. 5G rollout, connected devices, and rising customer expectations have exponentially increased network complexity. Legacy operational models, siloed data, bespoke analytics, and manual interventions are driving unsustainable OpEx growth. CSPs struggle with slow MTTR, alert fatigue, and reactive

The Solution

Google Cloud offers an opinionated, comprehensive framework for Autonomous Network Operations. This framework provides a strategic path to transform network operations from a reactive, human-driven model to a proactive, AI-powered autonomous state. The solution is built on three pillars: a unified data foundation known as the **Network Digital Twin**, an **AI-driven intelligence**

The Outcome

By adopting the Google Cloud Autonomous Network Operations framework, CSPs can achieve significant, measurable business outcomes. These include a dramatic reduction in OpEx by automating manual tasks, a faster MTTR through automated root cause analysis, improved service reliability by predicting and preventing outages, and the acceleration of new

Disclaimer: In no way, shape, or form should the results presented in this document be construed as defining an SLA, SLI, SLO, or any other TLA. The authors' sole intent is to offer helpful examples to facilitate a deeper understanding of the subject matter.

network management, impacting service reliability and customer experience.

layer that generates predictive insights, and an **Agentic Framework** that translates insights into automated actions.

service innovation by creating a more agile and responsive network infrastructure.

The Telco Imperative: Moving from Reactive to Autonomous

The State of Network Operations: A Patchwork of Complexity

The operational environment for a typical Tier 1 telco often involves a diverse array of systems that have evolved over many years. This typically includes existing OSS/BSS platforms from various vendors, and a range of specialized analytics tools tailored for specific areas such as RAN/Core performance or IP transport. A key challenge is that critical data (encompassing fault, performance, topology, and ticket information) often remains isolated within these disparate systems, preventing a comprehensive, unified view of network health and customer experience.

This fragmentation leads directly to two major operational pain points that erode profitability and customer trust:

1. **Reactive, Manual Troubleshooting:** When a service degradation occurs, troubleshooting becomes a slow, manual process of correlation. Skilled engineers are forced to become digital archaeologists, hunting for clues across disparate systems and dashboards. This inefficient process results in severe "alert fatigue," where critical signals are lost in the noise of thousands of uncorrelated alarms, and ultimately leads to extended MTTR.
2. **The Limitations of Batch Processing:** Many existing analytics platforms rely on batch processing, with data feeds that are hours or even a full day old. In a dynamic, real-time network environment, this latency renders insights obsolete upon arrival. By the time a potential issue is identified, the opportunity for proactive intervention is lost. This operational model is fundamentally reactive and cannot scale to meet the dynamic demands of 5G and edge computing.

Defining the Levels of Autonomy: A Maturity Model

The journey to full autonomy is an evolution, not an overnight revolution. We see this transformation happening across six distinct stages (as described by TMForum and shown in the table below), providing a clear maturity model for CSPs to benchmark their progress and plan their strategic roadmap.

P: Personnel **S**: Systems

Level definition	L0: Manual operation and maintenance	L1: Assisted operation and maintenance	L2: Partial autonomous network	L3: Conditional autonomous network	L4: High autonomous network	L5: Full autonomous network
Execution	P	P/S	S	S	S	S
Awareness	P	P	P/S	S	S	S
Analysis	P	P	P	P/S	S	S
Decision	P	P	P	P/S	S	S
Intent/Experience	P	P	P	P	P/S	S
Applicability	N/A		Select scenarios			All scenarios

The Business Case for Autonomous Network Operations Framework: Linking Technology to Value

The investment in the Autonomous Network Operations framework is directly tied to solving critical business challenges and unlocking new opportunities. The primary drivers include:

- **Operational Cost Reduction:** Automating manual tasks in the Network Operations Center (NOC) directly reduces headcount requirements and frees up highly skilled engineers to focus on service innovation rather than firefighting.

- **5G/Edge Monetization:** Guaranteeing stringent SLAs for new 5G and enterprise services, like network slicing and private 5G, is impossible with manual operations. The Autonomous Network Operations framework provides the assurance needed to monetize these advanced use cases.
- **Customer Experience and Retention:** Proactively predicting and preventing service-impacting outages is the most effective way to improve customer satisfaction and reduce churn in a highly competitive market.
- **Service Agility:** An autonomous, intent-driven network allows for the rapid deployment of new services and configurations, reducing time-to-market from months to days or even hours.

These drivers have been outlined in the Autonomous Network Operations framework: Unlock predictable and high-performing networks blog where we discuss the value proposition of Autonomous Network Operations framework on Google Cloud.

Google Cloud's Opinionated Autonomous Network Operations Framework: An Architectural Overview

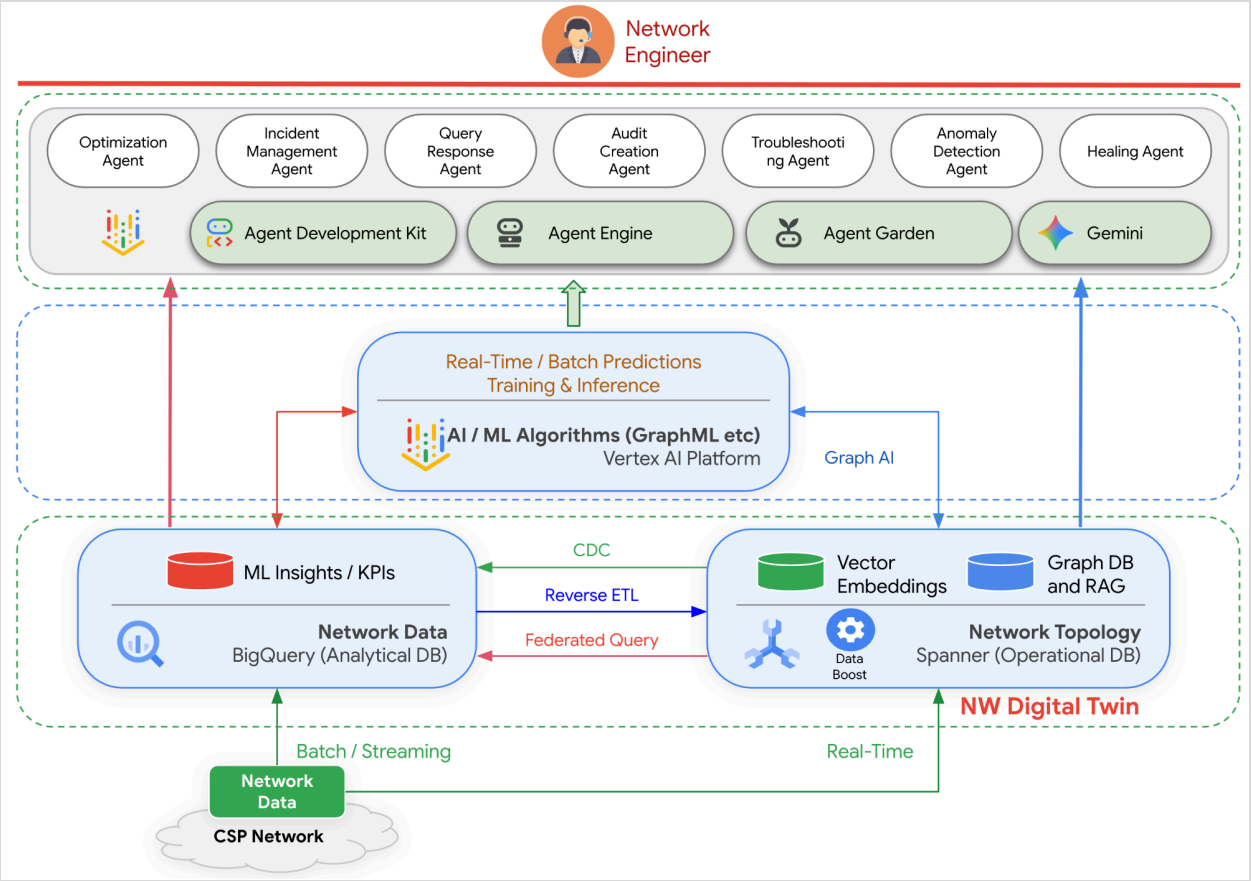
Core Principles

Our framework is built on a set of core, opinionated principles designed to address the fundamental challenges of legacy systems and deliver a truly modern operational environment.

- **Unified & Real-Time Data:** The framework's cornerstone is the elimination of data silos and the establishment of a single source of truth. By creating a unified data foundation that is updated in real-time, we provide a consistent, current, and holistic view of the entire network.
- **AI/ML at the Core:** We fundamentally shift the operational paradigm from simple data correlation, which is prone to error, to AI-driven causation and prediction. The goal is to move beyond observing what happened to understanding why it happened and predicting what will happen next.
- **Open & Interoperable:** The architecture is based on an interoperable stack of best-in-class, managed Google Cloud services. This provides maximum flexibility, avoids vendor lock-in, and allows for seamless integration with existing systems and

third-party tools via open APIs, ensuring that CSPs can leverage their existing investments.

The framework can be visualized as three interconnected layers that build upon one another to enable full autonomy, moving from data to insight to action.



Google Cloud Autonomous Network Operations High-Level Framework

Google Cloud Differentiators

Google Cloud is uniquely positioned to deliver this transformative framework due to three key differentiators:

- **Planet-Scale Infrastructure:** Our global network, serverless architecture, and distributed services like Spanner and BigQuery were designed from the ground up to handle data at an unprecedented scale, a prerequisite for managing a Tier 1 network.

- **Leadership in AI and ML:** Google is a world leader in AI research and development. This framework gives CSPs direct access to our most advanced technologies, including our proprietary GraphML and Gemini family of models, through the managed Vertex AI platform.
- **Fully Managed Services:** By providing core components as fully managed services, we abstract away the complexity of infrastructure management, allowing CSP teams to focus on building value, not managing VMs or database clusters.

The Foundation: Building the Network Digital Twin

The Concept: A Living, Temporal Model

The Network Digital Twin is the heart of the Autonomous Network Operations framework. It is not merely a static inventory or topology map, but a dynamic, multi-layer, real-time representation of the network. This includes:

- **Physical and Logical Topology:** From fiber optic cables and cell towers to virtual routers and network slices.
- **Operational State:** The real-time status of every element, including performance metrics, fault conditions, and configuration parameters.
- **Temporal Relationships:** Crucially, the Digital Twin is a temporal graph. It understands not just what the network looks like now, but what it looked like five minutes ago, five hours ago, or five days ago. This ability to query the network's state at any point in time is fundamental to performing accurate root cause analysis and understanding the evolution of network events.

The Unified Data Layer

We achieve this unified, high-performance view by leveraging two purpose-built, serverless database services to handle the distinct OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing) workloads of network operations.

- **Google Cloud Spanner (The Operational DB for OLTP):** Spanner serves as the foundation for the live, operational "System of Insight." It is designed to store and

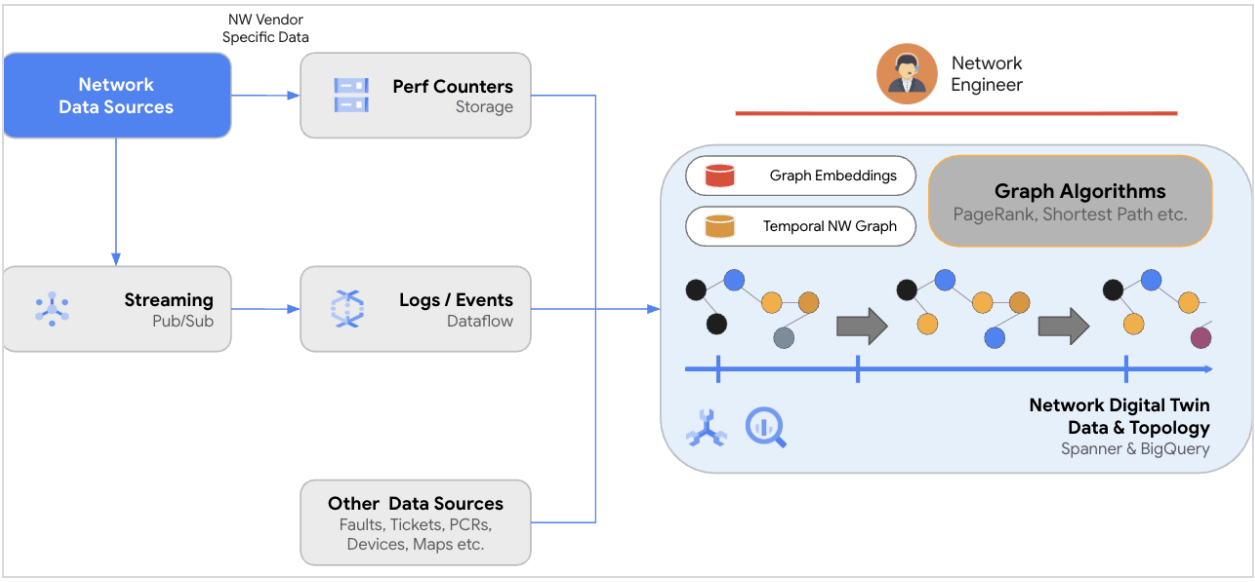
serve the live network graph, its complex topology, and its real-time state data. As the industry's only globally distributed, strongly consistent database with up to 99.999% availability, it provides the extreme reliability and low-latency query performance required for mission-critical operational use cases. Its native support for graph-like schemas and its ability to scale horizontally without downtime make it the ideal home for the live Digital Twin. A typical schema would model network elements (routers, switches, cell sites) as nodes and their physical or logical connections as edges, with properties on each representing their real-time state.

- **BigQuery (The Analytical DB for OLAP):** BigQuery acts as the consolidated analytical data warehouse. It is the destination for massive volumes of historical data, including performance counters (PM), logs, flow data, and telemetry which often amount to hundreds of terabytes per day for a Tier 1 operator. Its serverless, disaggregated architecture separates storage and compute, allowing for incredibly fast, complex analytical queries without provisioning any infrastructure. This enables deep, historical analysis, long-term trend identification, and the training of large-scale ML models without impacting the performance of the operational Spanner instance.

High-Throughput Data Ingestion at Telco Scale

The Digital Twin must be fed with a torrent of data from thousands of network sources. We utilize a serverless, highly scalable data ingestion architecture designed for this challenge.

- **Cloud Dataflow & Pub/Sub:** This combination creates robust, resilient, and massively scalable data pipelines. Pub/Sub provides a global, at-least-once delivery messaging service to ingest streaming data from thousands of sources in real-time. Dataflow offers a unified stream and batch data processing service to transform, enrich, validate, and load this data into Spanner and BigQuery. By using Dataflow templates, CSPs can standardize their ingestion logic, and the service's autoscaling capabilities ensure that the pipelines can handle extreme fluctuations in network data volumes without manual intervention.



Network Digital Twin Data Flow

Building the Temporal Network Graph: From Data to Model

Creating an effective Digital Twin requires a deliberate approach to modeling the network as a temporal graph.

- **Schema Definition:** The process begins with defining a unified data model, or schema, that can represent the heterogeneous elements of a telco network. This involves defining Node types (e.g., Router, CellSite, Port, CustomerService) and Edge types (e.g., CONNECTS_TO, DEPENDS_ON, CONTAINS). Each node and edge has properties that will be updated in real-time (e.g., port_status, traffic_utilization).
- **Entity and Relationship Extraction:** Dataflow jobs are configured to parse the incoming raw data (e.g., Syslog messages, SNMP traps, JSON telemetry) to extract these defined entities and the relationships between them. This is a critical ETL step that turns unstructured or semi-structured data into a structured graph format.
- **Modeling the Temporal Aspect:** To capture the network's evolution, every state change is stored as a new versioned entry or with a timestamp range. When an edge changes state (e.g., a port goes from "UP" to "DOWN"), the previous state's timestamp is closed, and a new state is recorded with a new start timestamp. This allows Spanner to support queries like, "Show me the exact topology and state of the network in the Ashburn data center yesterday at 2:15 PM."

The Intelligence Layer: Generating Actionable Insights with AI

Beyond Analytics: The Power of Graph AI

A network is not a set of independent data points; it is a graph of deeply interconnected entities. To truly understand network behavior, one must analyze it as such. Traditional ML models, which operate on tabular data, often fail to capture these complex relational dependencies.

- **Vertex AI & GraphML:** Our framework leverages Vertex AI, Google Cloud's unified ML platform, to train and deploy GraphML models directly on the temporal graph data stored in Spanner. GraphML models are purpose-built to learn from graph structures through a process called message passing, allowing nodes to share information with their neighbors. This enables them to perform tasks impossible for other models. Key use cases include:
    - **Advanced Root Cause Analysis:** When a fault occurs, GraphML can analyze the temporal subgraph leading up to the event. By tracking the propagation of abnormal signals (e.g., rising latency, packet drops) from node to node, it can pinpoint the likely root cause with high precision, rather than just identifying a wide range of correlated symptoms.
    - **"Blast Radius" Analysis:** Before a planned maintenance event, GraphML can simulate the impact of taking a device offline. By propagating the "outage" state across the graph, it can accurately identify all affected customer services, internal systems, and dependent network elements, preventing unforeseen consequences.
    - **Predictive Failure Detection:** GraphML excel at identifying subtle, system-wide patterns of degradation that are often precursors to a major outage. For example, a slight increase in latency across multiple, seemingly unrelated routers in a specific region might be a faint signal of a looming core network failure that traditional threshold-based monitoring would miss entirely.

Conversational Network Operations with Generative AI

The intelligence layer also aims to democratize access to complex network data, allowing engineers to interact with the system using natural language instead of complex query languages.

- **GraphRAG (Retrieval-Augmented Generation) Deep Dive:** We employ a powerful technique called GraphRAG. When an engineer asks a question, the system first uses an LLM to understand the user's intent and identify key entities (e.g., "fiber cut," "Ashburn," "affected customers"). It then queries the Network Digital Twin (Spanner and BigQuery) to retrieve a relevant, real-time subgraph and contextual data. This factual data is then passed along with the original question to the Gemini model. This "grounding" process ensures the model's response is based on factual, current network state, mitigating the risk of "hallucination" and providing accurate, trustworthy answers.
- **From Natural Language to Graph Query (NL2GQL):** For advanced use cases, the system can translate a natural language question directly into a formal Graph Query Language (GQL) statement. This allows technical users to leverage the full power of the underlying graph database without needing to write the queries by hand, dramatically improving their efficiency.

The Action Layer: The Agentic Framework for Automation

From Insights to Automated Actions

The final and most transformative step in achieving autonomy is to empower the system to act on the insights it generates. The Agentic Framework is a collection of specialized AI agents, each designed to perform a specific operational task. This creates a powerful, scalable, and auditable model for closed-loop automation, always with the critical inclusion of a "human-in-the-loop" for safety, oversight, and approval of any service-impacting changes.

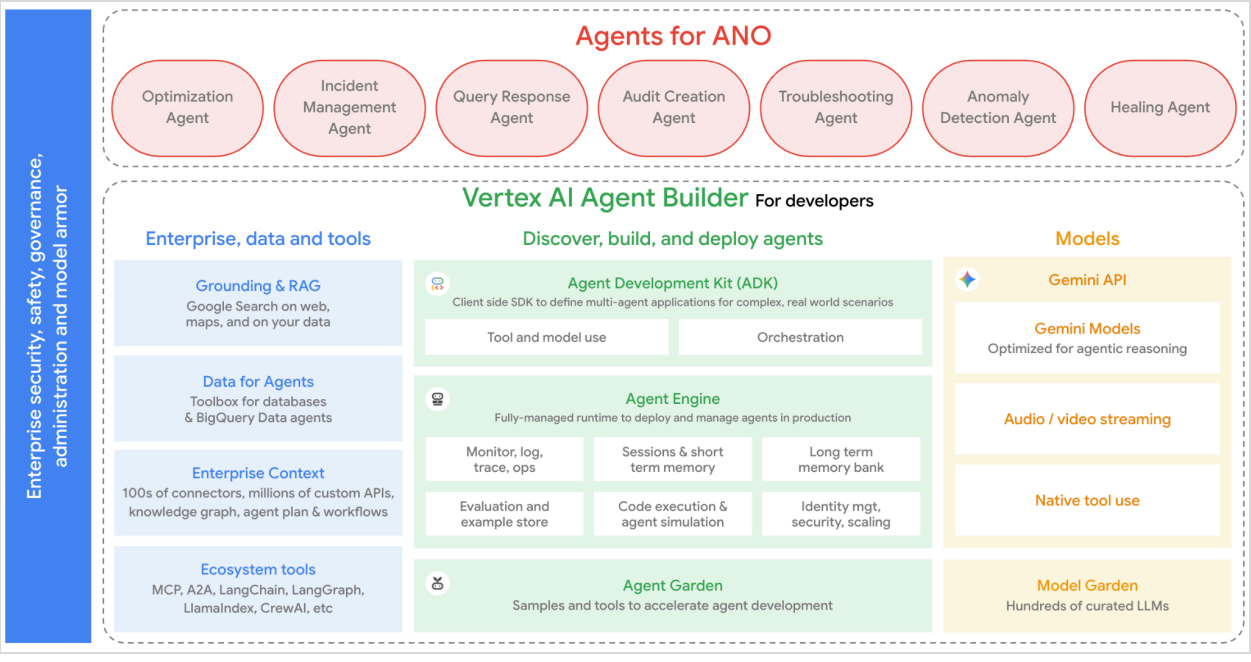
Building Autonomous Network Operations Framework Agents on Vertex AI

These agents are built and managed on Vertex AI. They function as intelligent orchestrators, consuming insights from the GraphML and RAG models and executing complex workflows. A suite of essential agents includes:

- **Anomaly Detection Agent:** Functions as the network's advanced nervous system. It subscribes to the real-time output from the GraphML prediction models, identifying

and flagging subtle deviations from normal network behavior and generating high-fidelity alerts with rich context.

- **Troubleshooting Agent:** Once an anomaly is detected, this agent automates the Tier 1/2 diagnostic sequence. It can execute pings and traceroutes, query the Digital Twin for device configurations and recent state changes, and correlate findings to isolate the problem domain, presenting a summary report in minutes.
- **Incident Management Agent:** Integrates with systems like ServiceNow to automate ticket creation. When an alert is validated, this agent can create an incident ticket, populate it with all the context from the Troubleshooting Agent (affected devices, potential root cause, customer impact), and assign it to the correct team.
- **Audit Creation Agent:** Periodically or on-demand, this agent queries the Digital Twin to retrieve current device configurations and compares them against predefined "golden configuration" templates or compliance rules. It generates reports identifying any configuration drift, flagging it for review.
- **Healing Agent:** This is the ultimate goal of automation. This agent takes a validated root cause and a proposed solution (e.g., from a library of Method of Procedures (MOPs) documents accessed via RAG). It first presents the action plan (including the exact commands to be run and the predicted impact) to a human engineer for approval. Once approved, it makes network api calls for corrective actions such as re-configuration, rerouting or restarting a network function.



The Autonomous Network Operations Agent Builder Framework

The Vertex AI Agent Builder

This entire agentic ecosystem is built using the **Vertex AI Agent Builder**. This comprehensive toolkit provides the necessary components to build, deploy, and manage these sophisticated agents at scale. It includes the Agent Development Kit (ADK) for defining complex, multi-agent workflows, and the Agent Engine for providing a fully-managed, production-grade runtime environment.

Implementation Path and PSO Partnership

A Phased Approach to Value Realization

We recommend a pragmatic, phased approach to adopting the Autonomous Network Operations framework, ensuring that value is delivered incrementally at each stage and that the organization can adapt to the new operational model.

1. **Foundational Data Consolidation:** The primary focus is on building the data ingestion pipelines and establishing the initial Network Digital Twin in Spanner and BigQuery for a specific, high-value network domain (e.g., the 5G core or a specific metro transport network).
2. **Initial Insights & Visibility:** Deploy initial analytics and dashboard use cases on the unified data. This phase delivers immediate value by providing engineers with a single pane of glass for network health, breaking down existing operational silos.
3. **Advanced AI-Driven Insights:** Begin developing and deploying the first GraphML models for use cases like predictive failure detection. Concurrently, deploy a GraphRAG-powered conversational agent to assist with common engineering queries.
4. **Closed-Loop Automation:** Incrementally deploy the Autonomous Network Operations framework agents, starting with non-intrusive agents like the Audit and Troubleshooting agents. Progress deliberately towards deploying the Healing Agent for specific, well-understood scenarios with strict human-in-the-loop controls.

The Role of Google Cloud PSO: Your Partner in Transformation

Google Cloud's Professional Services Organization (PSO) is a critical partner in this transformation. Our team of expert cloud data engineers and telco specialists works side-by-side with CSP teams to accelerate their journey. PSO's engagement model is collaborative and hands-on, providing:

- **Collaborative Workshops:** To define business goals and pinpoint the highest-value use cases.
- **Proof-of-Concept Development:** To rapidly prototype and validate the architecture with real customer data.
- **Co-Development and Implementation:** To build the foundational platform and initial use cases alongside the customer's team, ensuring knowledge transfer.
- **Tools and automation:** To accelerate deployments with data and graph migration / validation tools.
- **Best Practices and Governance:** To help establish a Network AI Center of Excellence (CoE) that can scale the Autonomous Network Operations framework across the entire organization.

Conclusion: The Future is Autonomous

The transition to Autonomous Network Operations is no longer a distant vision; it is a strategic necessity for survival and growth in the modern telecommunications landscape. The exponential growth in network scale and complexity demands a fundamental shift away from the manual, reactive, and siloed operations of the past. The Google Cloud Autonomous Network Operations framework (built upon a resilient foundation of a unified Network Digital Twin, an unparalleled AI-driven intelligence layer, and an extensible Agentic Framework) provides the technology, architecture, and expertise to make this transition a reality. By partnering with Google Cloud, CSPs can build a truly self-healing, self-optimizing network that not only lowers operational costs but also drives new revenue, accelerates innovation, and delivers an exceptional, reliable experience for their customers.

Call to Action

- Engage with your Google Cloud account team to learn more about the Autonomous Network Operations framework and initiate the adoption process.
- Consult your Google Cloud Professional Services team for assistance with planning and implementation.
- Explore the provided Colab notebook to begin building your Network Digital Twin on Spanner Graph.
- Checkout the starter RAN agent, a PoV accelerator for intelligent network operations, offers RAN KPI monitoring, anomaly detection, and root cause analysis. It is currently being open-sourced. Contact your Google Cloud account team for early access.

Appendix

Glossary of Terms:

- **CSP:** Communication Service Provider
- **MTTR:** Mean Time To Resolution
- **OLAP:** Online Analytical Processing
- **OLTP:** Online Transaction Processing
- **OpEx:** Operational Expenditure
- **RAG:** Retrieval-Augmented Generation
- **Digital Twin:** A dynamic, virtual representation of a physical network and its real-time state.

Product Links:

- Google Cloud Spanner
- Google Cloud BigQuery
- Google Cloud Vertex AI
- Google Cloud Dataflow
- Google Cloud Pub/Sub

Related Case Studies:

- Bell Canada launches AI-powered network operations solution built on Google Cloud
- Deutsche Telekom and Google Cloud Partner on Agentic AI for Autonomous Networks