LEARNING MADE EASY

Cloudera® Special Edition

Automating Telco Networks with AI

for dummies®

A Wiley Brand

- Build the foundation for network automation
- Learn about GenAI and agentic AI for telcos
- Make the shift to intent-based networking

Brought to you by

CLOUDERA

Jeremiah Morrow

About Cloudera

Cloudera is the only hybrid data and AI platform company that large organizations trust to bring AI to their data anywhere it lives. Unlike other providers, Cloudera delivers a consistent cloud experience that converges public clouds, on-prem data centers, and the edge, leveraging a proven open-source foundation. As the pioneer in big data, Cloudera empowers businesses to apply AI and assert control over 100% of their data, in all forms, improving security, governance, and real-time and predictive insights. The world's largest brands across all industries rely on Cloudera to transform decision-making and ultimately boost bottom lines, safeguard against threats, and even save lives. Learn more at cloudera.com.

Automating Telco Networks with AI

Cloudera® Special Edition

by Jeremiah Morrow

for dummies®
A Wiley Brand

Automating Telco Networks with AI For Dummies®, Cloudera® Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-394-45223-1 (pbk); ISBN 978-1-394-45224-8 (ebk); ISBN 978-1-394-45225-5 (ebk)

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Development Editor/Project Manager: Rebecca Senninger
Acquisition Editor: Traci Martin
Editorial Manager: Rev Mengle
Business Development Representative: Jeremith Coward
Production Editor: Tamilmani Varadharaj
Cloudera Team: Anthony Behan, Athul Prasad, Laura Blewitt, Wim Stoop, Claire Dominy, Jeff Healey

Table of Contents

Introduction

Telecommunications service providers are under operational, business, and competitive pressure to provide an exceptional customer experience while drastically reducing the cost of service delivery. The network is one of the industry's greatest expenditures, and network automation with artificial intelligence (AI) and machine learning (ML) provides the opportunity to both improve service and optimize costs. But automation efforts have stagnated, largely due to the volume, variety, and velocity of network data and the realities of data architectures.

About This Book

In this book, I explore the forces driving a renewed focus on network automation and some common barriers to success (Chapter 1). Then I discuss how operators can make progress with automation by starting with a solid foundation of network data (Chapter 2), implementing machine learning, generative AI, and agentic AI on that foundation (Chapter 3), and progressing your AI efforts to intent-based networking (Chapter 4). I'll look at techniques, such as bringing AI to the data, and what that means. By the end of this book you will have a roadmap for achieving network automation with AI.

Icons Used in This Book

Throughout the book, I use icons to indicate special information. Here's a guide to what those icons mean.

TIP

The Tip icon indicates information that you can apply to your own projects to make them work better. Tips can save you time and help you avoid frustration.

REMEMBER

The Remember icon highlights information that's worth retaining after you put down this book.

The Technical Stuff icon gives you detailed information related to a particular topic. Information marked by this icon isn't necessary for getting the job done, but it provides depth and interest.

Beyond the Book

This book introduces you to telco network automation and shows you how you can make progress on your automation journey with AI. The following resources will augment what you learn in this book:

- Video: How Data, AI, and Automation are Reshaping Telco Operations: `https://inform.tmforum.org/videos/how-data-ai-and-automation-are-reshaping-telco-operations`.
- Video: Three Ways Telcos Use Data & AI to Monetize: `https://www.youtube.com/watch?v=Dw7Aep70v7s`.
- Blog: The AI-Powered Data Lakehouse: `https://www.cloudera.com/blog/business/the-future-delivered-today-the-ai-powered-data-lakehouse.html`.

IN THIS CHAPTER

- The economic factors driving automation
- Making decisions at the speed of network data
- The barriers to automation

Chapter 1
Why Automation Matters More Now

Network operators have been pursuing automation since the early days of mechanical switchboards. However, for a variety of economic and technological reasons, automation matters more now than ever for telcos that need to increase shareholder value and meet customer expectations. Automating the network has the potential to optimize network operations and lower the cost of service delivery, push its performance limits, and achieve continuous service assurance and proactive customer support.

This chapter explores the trends driving automation initiatives, the role network automation plays in addressing the industry's challenges, and some of the common reasons why automation attempts fail.

The Economic Realities in Telco

The past several years in telecommunications (telco) have been defined by increasingly constricted margins. In 2024, worldwide revenues grew only 2 percent, according to a McKinsey report.

While revenues have remained generally flat, network throughput has increased significantly due to a drastic increase in video and streaming content, video conferencing and video calling, and, most recently, the proliferation of artificial intelligence (AI) and generative AI.

Subscriber growth has also stagnated, as markets around the globe are saturated. As a result, telcos are focusing on churn reduction and customer lifetime value (CLTV) growth as a means of protecting and growing revenue.

Telcos are also struggling underneath the weight of a burgeoning and aging network infrastructure. The network represents the largest capital expenditure for operators, and it has gotten more expensive to build and maintain. While most subscribers are leveraging 5G services, telcos are still managing infrastructure for 3G and 4G while starting to build for 6G. Maintaining the physical network infrastructure is putting significant strain on telcos.

Network automation has the potential to increase margins by reducing the cost of network operations and optimizing service delivery.

The Need for Speed: 5G and Beyond

The advancement of networking technology has made network automation a necessity. With potential speeds up to 200 times faster than its 4G predecessor and with near-real-time latency capabilities, 5G networks were supposed to transform our world, enabling use cases such as driverless cars, remote surgery, and augmented reality.

Despite its potential, 5G has largely failed to live up to the hype. One of the primary reasons for its lack of success is that network management decisions must be made faster than human beings can act.

If basic delivery of 5G service is difficult, the more transformative innovations of 5G, such as network slicing, are virtually impossible without automation. With each slice of the network possessing its own bandwidth and security requirements, it can quickly overwhelm human administrators.

As the industry moves toward 6G and the proliferation of more internet of things (IoT) devices, the volume of data makes manual oversight or classical heuristic-based automation solutions impossible. Without AI-driven automation to handle *closed-loop orchestration,* where the network senses and resolves issues in real time, advancements in networking infrastructure will continue to underwhelm. To unlock the true value of modern connectivity, we must shift from manual to automated (Chapter 3) and then to autonomous systems (Chapter 4).

Why Automation Attempts Fall Short

If automation is so important for modern telco operations, why are so many operators stuck in manual mode? The industry has been talking about *self-healing networks,* where issues are proactively detected and resolved autonomously to ensure continuous service, for a decade, yet the reality is that most network operations centers are still trying to move from reactive to proactive service assurance.

Automation efforts often fall short because of three fundamental barriers that legacy systems fail to overcome.

Scale and volume

The sheer amount of data generated by a 5G core is staggering: Telcos are monitoring millions of cells, virtualized functions, and billions of IoT sensors. Some telcos can ingest and process as much as six petabytes of data every single day. To put six petabytes of network data into perspective, if a single event record — phone calls, text messages, data sessions, cell handoffs, power records, faults, alarms — is around 200KB of data, that means 33 billion things are happening on the network every single day! Very few systems can handle that volume.

Data variety

In a perfect world, every piece of hardware in a network would speak the same language. In an actual environment, however, telcos are managing siloed legacy 3G/4G hardware, new 5G hardware from multiple vendors, and layers of on-premises and cloud-native software.

Data comes in hundreds of different formats, including structured, semi-structured, and unstructured data. Most automation attempts fall short because they rely on rigid tools that struggle to make sense of all these different data sources.

The latency gap

The latency gap is perhaps the most significant hurdle. For AI to automate a network, it needs access to data. Traditionally, high-compute AI models live in a centralized cloud environment. However, the data and network issues live at the edge.

If network data must travel from a local cell site to a centralized cloud or data center, get processed by an AI model, and then travel back as a command, the real-time window has passed. This backhaul latency makes cloud-only AI ineffective for mission-critical automation. Most current attempts fail because they can't get the intelligence close enough to real time to make a difference before a customer experiences an impact to their service.

Additionally, traditional monitoring tools were designed to collect data at five-minute intervals. In a high-speed 5G environment, at five-minute intervals, most network data is stale and has lost a significant amount of its value to the network operations team.

Despite these challenges, telcos must make progress on network automation to remain competitive and relevant in their markets.

IN THIS CHAPTER

- Shifting from batch to data in motion
- Using a data lakehouse
- Creating a unified data fabric
- Building edge-to-AI architecture

Chapter 2
Wrangling Network Data at Scale

At its core, network automation is a big data challenge. Most network data resides in on-premises environments, while the majority of AI/ML development and training occurs in cloud environments. The sheer volume, variety, and velocity of network data means moving it all to the cloud to do data science is often technically challenging and cost prohibitive.

To achieve network automation, the first step is for telcos to get a handle on their network data by addressing data architecture. Telcos have long been hampered by legacy, distributed systems that lack interoperability. This architecture has slowed innovation and made it more difficult to get real-time visibility into the network.

Telcos need access to the myriad data sources within the network and enable real-time ingestion, processing, and analysis at the edge. Then, they need to send pertinent network data back to the core for macro trend analysis and data science. Underneath all this data and AI work, they need to implement a consistent framework for security and governance to ensure compliance with regional and national data privacy and data sovereignty regulations and protect their customer data from cyberattacks.

These capabilities create the foundation of trusted data that is critical for the automation strategies I discuss in the upcoming chapters to succeed. Let's get started!

Creating a Data in Motion Pipeline

Traditionally, telcos have relied on batch ingestion and processing for network data. *Batch*, like its name implies, involves collecting data in chunks at certain intervals. These intervals can be extremely short, a matter of seconds (called *micro-batch processing*), but many legacy systems work with much longer intervals.

Whether telcos are relying on traditional batch or micro-batch processing, both are too slow and cumbersome to support automation, as shown in Figure 2-1. For use cases like anomaly detection and remediation, for example, telemetry data that could indicate a network issue loses more value every second it remains undetected.

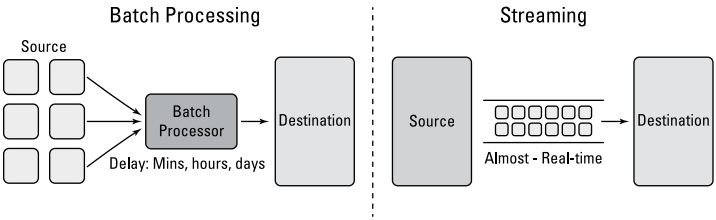


FIGURE 2-1: Batch versus streaming ingestion.

For the purposes of network automation, telcos implement a system called *data in motion*, which combines streaming ingestion, stream processing, and streaming analysis of data. If it's deployed at the edge, closest to the data's origin, telcos can get to near-real-time analysis and reduce the latency to milliseconds.

Data in motion leverages three systems:

- **Apache Kafka:** Kafka stores and distributes streaming messages. It ensures that data from thousands of cell towers get to the right applications without getting lost in transit.
- **Apache NiFi and MiNiFi:** NiFi helps you visually design the flow of information from the edge to the core. While NiFi

manages complex data routing at the data center, MiNiFi is a lightweight version that lives on edge devices to filter data at the source.

- **Apache Flink:** Flink enables analysis of data in motion, enabling telcos to spot patterns or anomalies in milliseconds. Instead of waiting for data to land in the database, Flink processes it in flight, so you can trigger automation scripts the moment a problem occurs.

Why the weird names? Most of the tools discussed in this chapter originated in the open-source community to solve big data problems at companies like LinkedIn, the National Security Agency, Netflix, and more. The Apache Foundation manages these open-source projects, and many vendors not only offer commercial, supported versions of these tools; they also actively participate in the continuous development and improvement of the tools.

Together, these tools create a data-in-motion pipeline that replaces batch systems and gives telcos access to data as close to the source as possible.

MiNiFi, the lightweight version of NiFi, is one of the most critical components of the data in motion architecture for telcos. When deployed on edge devices and used to preprocess data, MiNiFi supports intelligent ingestion by filtering a significant amount of the noise before transmitting it back to the core. Think about it: Telcos produce multiple petabytes of network data every single day. The resources required to move that much data from the edge to a data center or cloud every single day would be astronomical.

By preprocessing and filtering data at the edge, telcos have access to all their network data but transmit and store only the relevant data.

Storing Data in a Lakehouse

Like many other industries that have been around for a while, telcos have relied on traditional data warehouse platforms for most of their analytics and reporting workloads. While these systems have some benefits in terms of technological maturity and reliability, they have struggled to meet the needs of modern data volumes, especially in environments like telco networks.

Many telcos turned to data lakes, which worked great as cheap and efficient means of storing large volumes of structured and unstructured data. While these systems solved an immediate need, without the data management capabilities of the data warehouse, they quickly turned into "data swamps" and it was impossible to scale analytics and AI.

The solution to this challenge is an open data lakehouse, which combines the flexibility and scalability of data lake storage with the reliability and data management capabilities of the data warehouse.

REMEMBER

While advancements in AI have been grabbing the headlines, the maturation of the open data lakehouse is the real network automation enabler.

As shown in Figure 2-2, the data lakehouse consists of four layers that build on top of each other:

- **Storage:** Cloud or on-premises object storage is optimal. It can handle structured and unstructured data.
- **File format:** Data can be ingested in a variety of file formats. The most popular are JSON, Parquet, and ORC.
- **Table format:** The table format is how data is organized for analytics. It provides data warehouse management and governance capabilities directly on data lake storage.
- **Query engines:** Query engines are how users, from data engineers to business analysts, interact with data. Different engines are ideal for different workloads, and one of the biggest benefits of the open data lakehouse is multi-engine flexibility: the ability to use the best tool for each job.

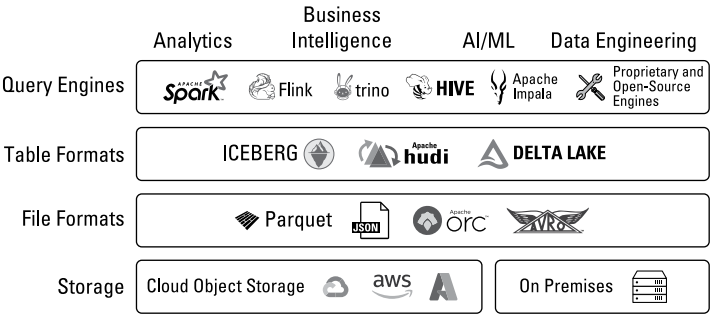


FIGURE 2-2: The open data lakehouse.

TECHNICAL STUFF

Apache Iceberg is an open table format for data lakes originally developed by Netflix and designed for high-performance analytics, simplified data management, and interoperability of large volumes of data. It's quickly becoming the de facto open table format for data lakes, with virtually every prominent data vendor supporting it and even contributing to the project.

The open data lakehouse can reside in an on-premises data center, in a public cloud environment, or distributed across environments in a hybrid architecture.

There are benefits to running different workloads, and even different parts of a workload, on specific infrastructure based on their requirements. AI model training, for example, is an intermittent workload that is most efficient in the cloud, where data scientists can spin compute up and down as needed. By contrast, AI inference is most efficient on premises where compute is more predictable, and where latency requirements might be more stringent.

TIP

The most important architectural consideration is to choose a data lakehouse platform with hybrid deployment capabilities that can run in the cloud or on premises.

Once data lands in the open data lakehouse, you can leverage it for a wide range of use cases:

- **Macro trend analysis:** Use Apache Flink to analyze data from edge sources and perform trend analysis across the entire network.
- **Dashboarding and reporting:** Deliver network health dashboards to customer service, field maintenance, the network operations center, IT, information security teams, and executive stakeholders.
- **Digital twins:** Construct a like-for-like digital replica of your network as a sandbox for network planning and optimization, scenario simulation, predictive maintenance, and physical infrastructure management.
- **Data science:** Build AI and ML models leveraging historical data from the data lake and push those models to the edge to run against real-time network data.

Building a Strong Unified Data Fabric

In a regulated industry like telco, security and governance of data is crucial. Telcos doing business in different regions and countries must adhere to a variety of data privacy laws or risk hefty fines. Additionally, many countries are passing data and AI sovereignty laws, forcing telcos to keep data regionally isolated.

A *unified data fabric* is the key to collecting, managing, and consuming data in a compliant way across this distributed architecture. It ensures consistent security, governance, and quality from the edge to the core. A unified data fabric is made up of three components:

- **Metadata management:** Operational and business metadata are important for data discovery, access, security and governance, and consumption, and it's at the core of a data fabric. Metadata ensures consistency across the entire data estate.
- **End-to-end lineage:** Lineage is critical for understanding where data originated, how it has changed, and how it is being used. It's also one of the most difficult elements of governance, especially in environments as complex as telco networks.
- **Security and governance:** Fine-grained access controls and enterprise-grade encryption of data at rest and in transit ensure data is protected at all times. For consistency, policies should travel with the data, so data teams only write them once, and they're enforced everywhere.

With a unified data fabric in place, telcos have the control they need to ensure compliance with regulatory requirements and data privacy laws, moving only necessary data and keeping it secure.

Putting It All Together: Edge to AI

The components we've discussed throughout this chapter are the building blocks for an edge-to-AI architecture, as shown in Figure 2-3. The pieces all fit together with data flowing through the system:

- **Streaming data at the edge:** Data is collected from edge devices, preprocessed, and analyzed in streams. MiNiFi sends pertinent data back to the core.
- **The unified data fabric:** A data fabric that ensures end-to-end consistency, security, governance, data quality, and visibility into the data architecture weaves the entire platform together.
- **The network data lakehouse:** Data lands in the data lakehouse, where it is processed, standardized, enriched, and made available for reporting and data science.
- **Data science:** Data scientists update models using fresh data and push those models back to the edge for use on streaming data.
- **Reporting and dashboarding:** The network data lakehouse provides a near-real-time view of network health via dashboards and reports for various departments and stakeholders.

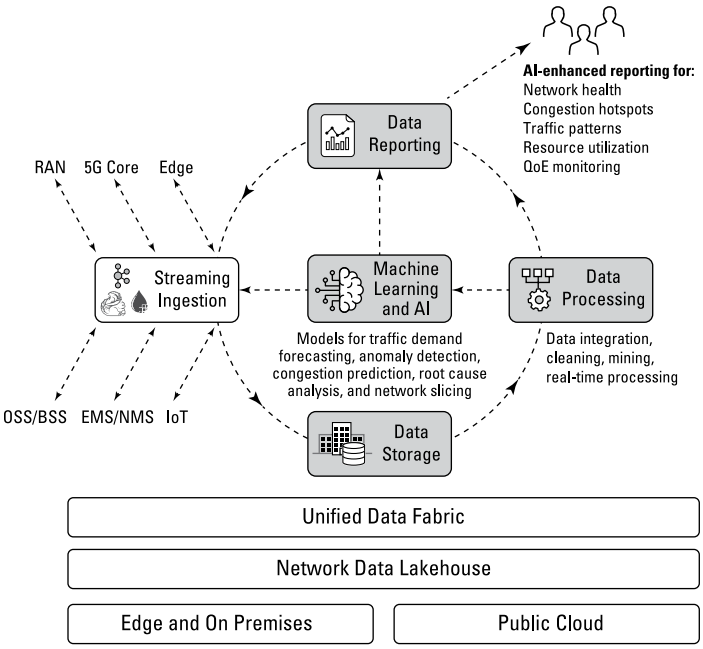


FIGURE 2-3: Edge-to-AI architecture in telecommunications.

The result is a system that provides real-time network intelligence and automation at the edge, visibility into the entire system, and the foundation for AI-powered innovation and optimization.

IN THIS CHAPTER

- Starting with machine learning
- Progressing to generative AI
- Going autonomous with agentic AI
- Giving AI context

Chapter 3
Machine Learning, Generative AI, & Agentic AI

Once you've built the foundation and you have access to a consistent, high-quality view of network data, it's time to start building the models that will automate the network.

AI technology is advancing at an unprecedented rate, and you have some decisions to make in terms of what types of AI make the most sense based on your use cases. This chapter explores the three most common and mature AI types, what types of analytic workloads they work best for, and recommends some high-value use cases to get started.

Machine Learning

Although AI is currently enjoying the limelight, many use cases don't require the compute resources required by generative and agentic AI models. Machine learning (ML) can still do some of the heavy lifting.

Machine learning is essentially a computer solving a statistical equation. It takes an algorithm, runs it against a large volume of data, and learns to spot patterns and anomalies. One classic example is training a machine learning algorithm by showing thousands of pictures of a dog. It can eventually learn the physical attributes that make a dog a dog, and after it's trained, it should be able to identify whether an animal shown in a picture is likely a dog. This kind of pattern repetition also plays a role in AI.

Pattern recognition is perhaps the largest component of network automation, and so machine learning is sufficient for many initial use cases. By running ML against streaming telemetry data at the edge and writing scripted responses, network engineering teams can enable the algorithm to detect anomalies in large volumes of fast-moving data and remediate many network issues in real time.

A couple of high-value use cases where machine learning excels include:

- **Anomaly detection:** Log data from network machines is massive in volume and velocity, and spotting anomalies is nearly impossible without the assistance of machine learning. Fortunately, machine learning excels at this use case, because it's all about pattern recognition. By training a model on historical network data and deploying it at the edge, it can flag anomalies for support staff and field maintenance and even run scripts to address some issues.

  For example, consider a Distributed Denial-of-Service (DDoS) attack, a common network threat where the attackers use a sudden increase in traffic to overload systems and take down the network. A machine learning algorithm trained to detect malicious requests from legitimate user traffic can detect the threat and apply filtering rules to allow legitimate traffic to pass through, stopping the attack without interrupting service.
- **Predictive maintenance:** The costs and complexities associated with cell site maintenance are increasing substantially, especially as operators continue to roll out 5G service.

  The hardware that makes up the network produces an enormous amount of data, and signals within that data can indicate a potential device failure is looming. Machine

learning algorithms trained on historical device data, sales and service records, and maintenance logs can run against log data from network hardware and flag a potential failure before it occurs.


TIP

While some of the physical assets of a telco network have no data associated with them, digital twins represent an opportunity to digitize physical assets and predict maintenance needs based on historical data.

By taking a proactive approach to network maintenance, telcos can reduce unplanned downtime, extend the life of their hardware assets, and optimize maintenance schedules.

Generative AI

Generative AI (GenAI) is trained on vast amounts of text to understand, summarize, and generate human-like language. Public models such as ChatGPT, Gemini, and Anthropic Claude are based on large language models (LLMs), meaning they are trained on nearly the entire public internet.


TECHNICAL STUFF

While LLMs are more prominent and well-known, small language models (SLMs), stripped-down versions of LLMs that are trained on specific network functions, are incredibly useful for telcos. Because of their reduced size and scope, they can run at the edge with significantly reduced latency.

Because they work best when a human being is in the loop, GenAI is at its best as a copilot or assistant:

- **Troubleshooting:** Like their ML counterparts, GenAI can review massive volumes of log data in near-real time, much faster than a human being. In terms of telemetry, this means they can detect anomalies in a fraction of the time it might take a human analyst. They can also review, synthesize, and even write documentation, and suggest fixes.
- **Customer support:** Whereas the chatbots of the past essentially followed a script, GenAI chatbots can understand intent. They can understand and respond to a customer's needs in a personalized way, improving outcomes and customer satisfaction. As a result, telcos can resolve more customer interactions without the need for human intervention.

Going Autonomous: Agentic AI

Agentic AI represents a massive opportunity for telcos: It can enable the shift from automated to autonomous.

Whereas ML is limited to pattern recognition and executing simple scripts based on if/then clauses, and GenAI is limited to providing and synthesizing information, agentic AI can act. It can use tools, run applications, and determine the next best action based on a given prompt.

Agents use chain-of-thought reasoning where they break a complex goal down into smaller tasks, choose the right tools for each task, and check their own work as they go.

Implementing AI agents that can execute important tasks within a network requires significant governance and oversight. Although the governance will always be necessary, human oversight in agentic workflows can sometimes be phased out over time. This approach to autonomy with a self-healing network agent can happen in phases:

- **Human in the loop:** The agent detects a power failure at a tower. It analyzes the logs and realizes a backup battery has kicked in. The AI sends an alert to a human technician's dashboard with a Fix button. The human reviews and approves, and the agent reroutes traffic.
- **Human on the loop:** The agent detects a power failure. It immediately reroutes traffic to nearby towers and schedules maintenance for the morning because it knows the battery will last 12 hours. It sends a notification to a human supervisor with override authority if the agent has taken the wrong action.
- **Human out of the loop:** The power goes out, and AI instantly optimizes the entire regional cluster, adjusts neighboring antennas to cover the dark spot, and determines when the grid will be restored based on the local utility company's notifications. A human reviews all AI actions in a report. The human's role shifts here from fixing the network to setting the goals for AI agents to achieve, which I discuss in the next chapter.

REMEMBER

Moving out of the loop requires a massive amount of trust in the data and in the AI. This is why the unified data fabric, and especially the end-to-end lineage from the previous chapter, is so important. If AI makes a decision, it needs to be auditable and explainable.

AI agents work best in scenarios where autonomy is paramount for success. A couple more agentic AI use cases can yield big benefits for telcos.

5G network slicing

5G network slicing is network virtualization on steroids. Within a massive 5G network pipe, telcos can segment traffic based on workload criticality, so traffic that requires significant or dedicated bandwidth have access to it. 5G network slicing was supposed to fundamentally transform telcos from network providers to innovation enablers, but it largely failed to live up to expectations, primarily because of a lack of automation.

When mission-critical workloads like remote surgery and driverless vehicles fail, there are major consequences. Sometimes, bandwidth allocation decisions must happen faster than a human can make them.

Network slicing agents can finally make this 5G promise a reality by anticipating and dynamically adjusting bandwidth allocation based on network traffic volume and workload criticality, ensuring dedicated, always-on bandwidth for the most important traffic.

5G network slicing can also give telcos the ability to sell dedicated bandwidth at higher premiums than normal traffic. Agentic AI can make those opportunities a reality.

Autonomous energy optimization

In their current state, cellular networks generally operate at full power 24/7, even when the vast majority of customers are not using the service. This represents a massive operating expenditure and impacts environmental, sustainability, and governance (ESG) goals, especially as AI demands greater levels of energy and water consumption.

Agentic AI can act as the smart power manager for the Radio Access Network (RAN). Instead of a timer that turns everything off at a certain time, an AI agent can monitor real-time traffic patterns, weather reports, and even local event schedules to determine traffic requirements. It can power down network equipment when demand drops and wake it up in milliseconds if there is a sudden spike in traffic to maintain service quality.

Energy is often the second largest operating cost for a telco after labor. Moving from a static power schedule to an agentic one can reduce energy consumption by 10-15 percent without impacting customer experience.

Ensuring AI Accuracy and Consistency with Context

Regular users of public AI tools know two things about LLMs: They make things up at an alarming rate, and they are probabilistic, not deterministic. Given the exact same prompts multiple times, you can generally expect different outputs each time.

Given these two realities, it's sometimes difficult to imagine trusting AI to run core business processes autonomously. That's why you need to ensure that any AI you use has the context it needs to deliver accurate and consistent outputs.

The good news is context starts with a foundation of trusted data, something I discuss in the previous chapter. Now, you can leverage one or several methods to ensure AI builds on top of that foundation.

REMEMBER

When choosing which method of AI model refinement — RAG or fine-tuning — is best, it's important to make decisions based on individual workload requirements. Each can excel in different scenarios, and the best telcos use both!

Retrieval-augmented generation

Retrieval-augmented generation (RAG) enables an AI model to run against your foundation of trusted data in real time, at the moment it receives a query.

Streaming RAG, built with the data-in-motion suite of tools, is ideal because it ensures that AI's outputs reflect the near-real-time state of your network. NiFi picks up the data and routes it to the AI's vector database, Kafka ensures AI receives millions of events in the exact order they happened, and Flink cleans and summarizes data in motion to reduce the data AI sees.

In a streaming RAG architecture, shown in Figure 3-1, the data-in-motion tools collect data in real time from the network. It enters a vector database, where it's available for AI to query against. When the LLM receives a query, it hits the vector database, as well as other data sources, and then it uses that data to generate a response. The user receives an answer to their query based on the most recent state of the network.

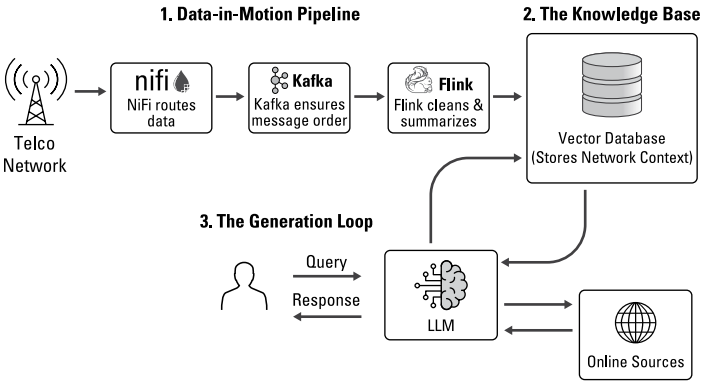


FIGURE 3-1: A streaming RAG architecture.

RAG has some real benefits for telcos. First, it provides real-time context, so in an environment where the network and equipment status change from moment to moment, AI will always reflect the current state. Second, it virtually eliminates hallucinations because AI must cite sources from your data. Streaming RAG is ideal for many network use cases that can't tolerate latency.

Fine-tuning

Fine-tuning is the process of taking a base AI model and making it specialized by showing it thousands of examples of exactly how you want it to behave. It's best for using AI in situations where technical or very specific knowledge is critical, such as understanding specific regulatory standards and requirements or writing specialized network code.

The major benefits of fine-tuning are that models are faster and more precise at specific tasks. In fact, with fine-tuning, SLMs can often outperform much larger, more expensive models.

Staying within a Private AI System

Telco is a highly regulated industry with significant security requirements, sensitive customer information, and valuable intellectual property. A lot of important data must remain behind the firewall, not exposed to the public internet.

Private AI is where the AI model lives entirely within a telco's controlled environment. You maintain full control over models, data, and infrastructure associated with AI. From a network perspective, private AI generally means data and AI remains on premises, where the network data lives. Private AI is critical for a few reasons:

- It enables data sovereignty, where customer data never crosses national borders.
- It means models and model development remain entirely in-house. If public AI models change, customers become responsible for regression testing everything built on an older version, which creates significant operational overhead that may not have been necessary. The older model may have been working just fine.
- It's much easier to audit and prove compliance to regulators and governments when data remains under your control.

IN THIS CHAPTER

- Transitioning to imperative goals
- Self-healing networks with a continuous assurance loop
- Using natural language to converse
- Putting up AI guardrails

Chapter 4
Closing the Loop with Intent-Based Networking

Once you introduce automation and autonomy into network operations, it's time to change how you manage the network. Intent-based networking is a shift from using rules-based systems to tell AI and ML how to do their jobs to setting high-level goals and letting AI decide how to execute against it.

In this chapter, I cover three important components of intent-based systems: declarative network protocols, the continuous assurance loop, and natural-language NetOps.

Shifting from Imperative to Declarative Goals

To understand intent-based networking, you first have to understand how the industry has been doing things for the last 30 or so years. Traditionally, network management has been imperative.

In an imperative model, you give the network specific instructions that the system executes. To reroute traffic, you write scripts, configure ports, and update the routing tables manually.

The primary transformation of IBN is shifting from an imperative to a declarative model. In a declarative model, you simply state your intent or goal. AI takes that business goal and translates it into the thousands of technical configurations needed across routers and switches. In an agentic system, AI can push those configurations out to the network automatically.

Providing AI-Powered Continuous Service Assurance

Like imperative network protocols, service assurance has evolved significantly in the last several years.

For decades, networking was entirely reactive. Network engineers set thresholds, and if those thresholds were exceeded, it sent an alert to a dashboard. The primary challenge with this system was that once the dashboard turned red, the user experience had already been impacted. This era was defined by long resolution times and manual root cause analysis.

Around 2020, telcos shifted toward proactive management, powered by early ML models and streaming telemetry. Instead of waiting for a network issue, systems could use pattern recognition and mathematical equations to identify the signals that might predict a potential disruption.

Continuous assurance is the next era of service assurance for telcos. It moves beyond simply knowing there is (or there might be) a problem to resolving it automatically, or *self-healing.*

In a continuous assurance model, the network operates in a perpetual cycle often called the Observe-Orient-Decide-Act (OODA) loop. Because you have provided a declarative goal, the system is constantly checking the actual state of the network against the declarative goal. Here is how the OODA loop works:

- **Observe:** The data-in-motion pipeline feeds real-time telemetry into the system.

- **Orient:** The AI compares this live data against data in the network data lakehouse to compare the actual state against the intended state.
- **Decide:** If there is a drift, agentic AI reasons out a solution.
- **Act:** The agent executes a change.

The loop closes when the system observes the results of its own action to ensure the fix actually worked.

With AI-powered continuous assurance, the focus for telcos can shift to Mean Time to Resolution (MTTR). With agentic AI, MTTR can drop from hours to milliseconds. In many cases, the network self-heals so quickly that the customer, and even the network operations center, never knew there was a problem in the first place. This is called *zero-touch resolution.*

Switching from Commands to Natural Language

Up until now, we've discussed how AI thinks and how data flows through the system. But how do humans actually interact with it? For decades, the only way to talk to the network was through a CLI, a black screen where users input code and precision is critical.

AI has given us a better way. Natural language NetOps changes the interface from code to conversation. By layering a GenAI copilot over the agents and the network data lakehouse, you can empower the network operations team to chat with their network using natural language.

A couple of technical components are critical for enabling natural language NetOps:

- **Lineage:** Metadata and lineage are critical for ensuring your copilot and network team are speaking the same language and understanding the network topology the same way.
- **The network data lakehouse:** Not everyone who needs to interact with the network data lakehouse knows how to write SQL. The copilot translates a natural-language prompt into a SQL query and runs it against the lakehouse.

- **Governance:** The copilot follows access control policies established with the unified data fabric to ensure the user has permissions to access the data.
- **Agentic AI:** Copilots can also be used to set intents. A user sets a declarative goal and AI agents execute against it.


TIP

Natural language NetOps enables junior technicians to perform more complex troubleshooting and senior engineers to focus on strategy instead of syntax errors.

Setting Guardrails: Creating a Safety Net for AI

Giving an AI agent the power to change network configurations is exciting, but it's also risky. Implement these guardrails to protect your network, your customers, and your business:

- **The policy layer:** Before Agentic AI can push a change, it must pass through a hard-coded policy engine. These are unbreakable rules set by humans.
- **The sandbox simulation:** Digital twins de-risk autonomous decisions made by AI, and this is the perfect situation to use it. Before activating a declarative intent in the real world, network operations teams can test it against a digital twin.
- **Explainable AI:** Although the system is autonomous, observability is still important. To understand why AI made certain decisions and to satisfy regulators and executives, network operations teams must be able to explain AI outputs. This is where lineage and governance across the entire system becomes critical.


TIP

Using time travel, a feature of Apache Iceberg, the open table format for the network data lakehouse, you can use metadata to recreate a view of Iceberg tables at a specific point in time. Now, you can actually see what the AI saw when it made a decision!

IN THIS CHAPTER

- Building a strong foundation of data
- Layering AI on top
- Adding processes and people skills

Chapter 5
Ten Steps to Building an Autonomous Network

This chapter summarizes the ten steps to building an autonomous network.

- **Audit your data silos.** Before you can automate the network, you need to know what you're working with. Identify where your network data lives and what systems it lives in. Bring all the data together, ideally with a data fabric that provides unified access, security, and governance.
- **Start at the edge.** Moving every petabyte of raw network data to the core is technically challenging and cost prohibitive. Use lightweight tools like Apache MiNiFi to pre-process and filter data at the source, ensuring only relevant signals travel across the network. This process ensures that only the signals that matter travel across the network, reducing costs and improving efficiency.
- **Build a network data lakehouse.** Use a data lakehouse with open table formats such as Apache Iceberg to store historical data. The lakehouse serves many functions, including model training, reporting and dashboarding, and digital twin operations. In an autonomous, intent-based

network, it enables AI to compare real-time telemetry against what a healthy network looks like.

- **Prioritize high-value use cases.** Don't try to automate the entire network on day one. Start with low-risk, high-reward projects such as energy optimization or DDoS detection.

  These use cases prove value to stakeholders and build the internal trust needed for bigger projects.

TIP

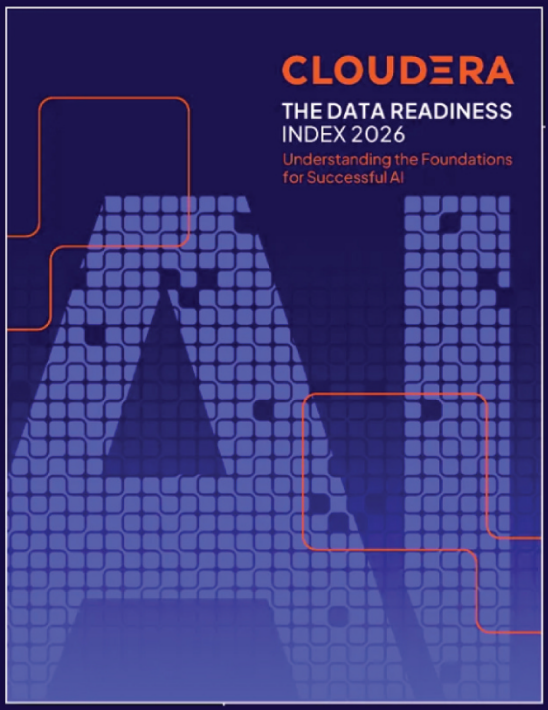
- **Choose the right tool for the job.** All the architectural and technology decisions you make give you unparalleled flexibility and adaptability in how you approach network operations from the edge to the core. Choose the right engine, AI, and model optimization for each use case.
- **Crawl, walk, and then run towards autonomy.** Autonomy doesn't happen overnight. Take a phased approach. Start with AI suggesting fixes with human approval before moving to AI acting and humans supervising. After you have built institutional confidence, you can move to full autonomy.
- **Shift from imperative to declarative goals.** Imperative scripts are rigid and prone to breaking. And in an autonomous system, they're unnecessary. Train teams to set intents, which are declarative goals that tell agentic AI what to achieve rather than how to do it.
- **Put AI guardrails in place.** Your autonomous network can't be a black box. Use lineage and governance tools to ensure actions taken by AI agents are auditable and explainable. Use digital twins as a sandbox to de-risk autonomous decision making. Finally, ensure critical policies are hard coded into a policy engine so AI can't break the network.
- **Invest in skills, not tools.** Shift your training focus away from manual CLI configuration to prompt engineering and data science. Think of your engineers less as mechanics and more as architects of an intelligent system. By building the muscles you need to be an AI-powered telco, you'll make the network a point of differentiation, and you'll attract and retain people who want to build and manage AI systems.
- **Iterate constantly.** The Observe-Orient-Decide-Act loop is critical for autonomous, intent-based networks, and for your business. Use fresh data from the network and insights from the network data lakehouse to constantly refine AI models.

CLOUDERA

THE DATA READINESS INDEX

Telcos Hold Highest Level of Fully Governed Data

Massive, distributed telco environments create complex data and high stakes. Maintaining performance is one of those stakes.



CLOUDERA

Read Cloudera's Data Readiness Index for more.

Fully automate your telco network

Automating your telco network to reduce operational costs and optimize service delivery is essential for future growth. But automation in telco is often harder than it sounds, even with modern advancements in AI. This book will guide you on your automation journey, starting with the foundational data architecture, exploring key AI and ML concepts and use cases, and finally transitioning from "automated" to "autonomous" with agentic AI and intent-based networking.

Inside...

- Learn why automation is necessary
- Get your network data in order
- Create a unified data fabric
- Start with machine learning
- Add generative AI capabilities
- Go autonomous with agentic AI
- Deliver continuous service assurance

CLOUDERA

Jeremiah Morrow has been writing about technology, data and AI, and digital transformation for more than a decade. As an industry product marketer, he specializes in making complex technical topics approachable. He lives in Annapolis, MD with his wife, a toddler, and their black lab, Coffee.

ISBN: 978-1-394-45223-1
Not For Resale

for dummies®
A Wiley Brand

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.