From Experiment to Enterprise: Building Scalable, Observable MLOps Systems on Google Cloud

By : Ancilia Dmello
Ford motor company
Conf42 Machine Learning 2026

The Production ML Challenge

Moving machine learning from experimentation into reliable, large-scale production remains one of the most critical challenges organisations face today. The promise of ML-driven insights often collides with harsh operational realities.

Infrastructure complexity, fragmented tooling, stringent governance requirements, and persistent model degradation create formidable barriers that prevent teams from realising sustained business value from their ML initiatives.

CORE BARRIERS

What's Stopping Enterprise ML?

Infrastructure Complexity

Managing distributed systems, orchestrating workflows, and maintaining compute resources across development and production environments

Fragmented Tooling

Stitching together disparate tools for training, deployment, monitoring, and governance without unified workflow integration

Model Degradation

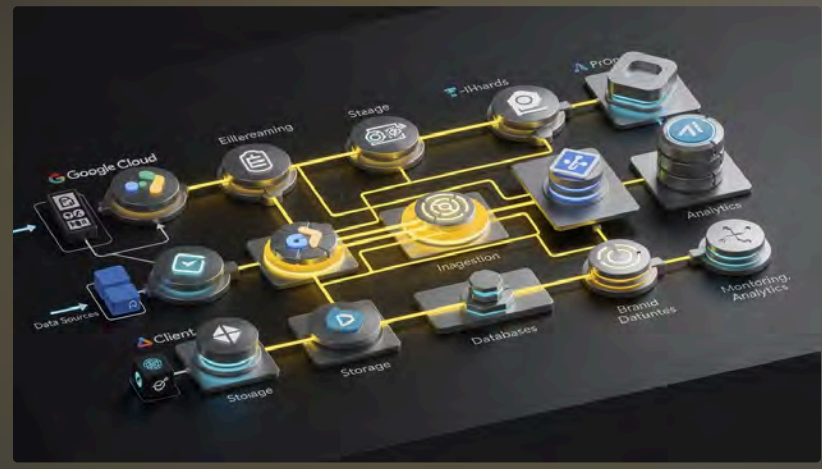
Data drift, concept drift, and feature skew silently erode model performance over time, impacting business outcomes

Governance Requirements

Meeting compliance mandates, ensuring auditability, and implementing human oversight whilst maintaining operational velocity

A Production-Grade MLOps Framework on GCP

This talk presents a cohesive, end-to-end system built on Google Cloud Platform that addresses these challenges head-on. By composing managed services into an integrated architecture, organisations can operationalise ML at scale whilst maintaining reliability, observability, and governance.



GCP SERVICE STACK

The Managed Services Foundation

- **Vertex AI**

  Unified ML platform for training, deploying, and managing models with built-in experiment tracking and pipeline orchestration

- **Cloud Run**

  Serverless compute for containerised model serving with automatic scaling and pay-per-use pricing

- **BigQuery**

  Serverless data warehouse for feature engineering, model training data, and prediction logging at petabyte scale

- **Dataflow**

  Managed stream and batch processing for real-time feature computation and data pipeline orchestration

- **Cloud Monitoring**

  Observability platform for metrics, logs, traces, and custom model health indicators across the entire system

Deployment Patterns: Architectural Trade-offs

Selecting the Right Approach

Different workloads demand different deployment strategies. Real-time inference prioritises sub-second latency for immediate decision-making. Batch processing optimises throughput for offline scoring. Streaming patterns balance both for continuous analysis.

Your choice depends on business constraints: latency requirements, prediction volume, cost tolerance, and operational complexity. Each pattern leverages GCP services differently to meet specific performance envelopes.

Production Observability: The Critical Foundation

Without comprehensive observability, production ML systems operate in darkness. Early detection of degradation whether from data drift, infrastructure issues, or prediction anomalies is essential for maintaining reliability and customer trust.

A robust observability strategy combines model-level and infrastructure-level telemetry to provide complete visibility into system health and performance.

MODEL MONITORING

Detecting Model Degradation

- **Data Drift Detection**

  Statistical tests comparing incoming feature distributions against training baseline to identify population shifts

- **Feature Skew Analysis**

  Monitoring discrepancies between training and serving feature values that indicate pipeline inconsistencies

- **Prediction Drift Tracking**

  Detecting shifts in model output distributions that signal changing behaviour patterns or model staleness

- **Performance Degradation**

  Tracking accuracy, precision, recall, and business KPIs against ground truth when labels become available

INFRASTRUCTURE TELEMETRY

System Health Visibility

- **Target Availability**

  Monitor uptime, error rates, and service health across all inference endpoints

- **Latency Thresholds**

  Track P50, P95, P99 response times to ensure SLA compliance for real-time workloads

- **Autoscaling Efficiency**

  Observe scaling behaviour, resource utilisation, and cost optimisation opportunities

Infrastructure-level observability complements model monitoring by surfacing operational issues latency spikes, resource exhaustion, scaling bottlenecks that degrade user experience regardless of model quality. Cloud Monitoring provides unified dashboards, alerting, and integration with incident management systems.

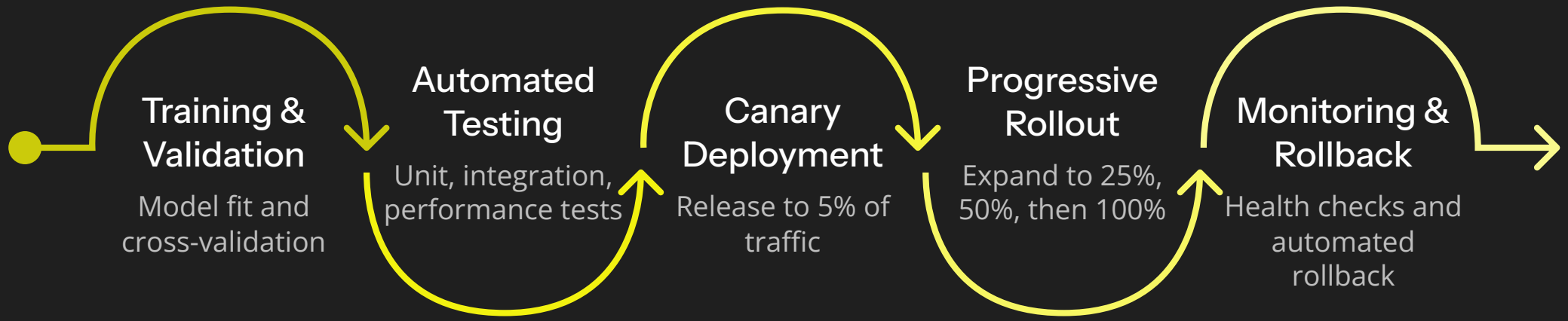
CI/CD-Driven Model Deployment

Deployment Best Practices

Production ML requires the same rigour as software engineering: versioning, automated testing, progressive rollouts, and quick rollback capabilities. A mature CI/CD pipeline ensures every model deployment is reproducible, auditable, and reversible.

This approach balances operational velocity with risk management, enabling teams to ship improvements confidently whilst maintaining production stability.

Deployment Pipeline Architecture



This progressive deployment strategy minimises risk by exposing new models to increasing traffic volumes whilst continuously monitoring performance. Automated health checks at each stage enable rapid rollback if degradation is detected, protecting customer experience.

GOVERNANCE & COMPLIANCE

Automated Retraining with Human Oversight

Balancing Automation and Control

Automated retraining pipelines keep models fresh as data distributions evolve. However, regulated environments require human-in-the-loop approval gates to satisfy compliance mandates and manage operational risk.

This framework implements scheduled retraining triggers, automated validation against hold-out datasets, and approval workflows that route high-risk changes to domain experts whilst auto-approving low-risk updates. Comprehensive audit logs satisfy regulatory requirements.

GOVERNANCE CONTROLS

Auditability, Versioning & Rollback

- **Model Versioning**

  Immutable model artefacts with semantic versioning linked to training data, code, and hyperparameters for complete reproducibility

- **Audit Logging**

  Comprehensive logs capturing who deployed what model, when, with what approval, and what the outcome was

- **Instant Rollback**

  One-click rollback to previous model versions with automated traffic shifting and health verification

- **Policy Enforcement**

  Automated checks ensuring models meet fairness, privacy, and security requirements before production deployment

Your Practical Blueprint for Production ML

This framework provides a comprehensive, battle-tested approach to building resilient, maintainable, and observable machine learning systems on Google Cloud. By combining managed services with robust observability, governance, and deployment practices, organisations can finally bridge the gap between experimental ML and reliable enterprise production.

Thank You!
Questions and Discussion.?

- Ancilia Dmello || Ford motor company || Conf42 Machine Learning 2026