

Connecting Your Laptop to Present



KubeCon



CloudNativeCon

Europe 2026

If you plan to use speaker notes, please follow the steps below to ensure your laptop is set up correctly. This will allow you to view your notes while the audience sees only your slides.

Before Connecting

- Close all unnecessary apps & tabs
- Turn off notifications
- Disable Night Mode / TrueTone
- Don't enter full-screen presentation mode yet

At the podium

- Plug in the HDMI cable
- If needed, use your USB-C to HDMI adapter
- On Mac, click "Allow" if prompted

USE "EXTENDED DISPLAY" MODE (NOT MIRRORED)

Using MAC

- Go to System Settings > Displays
- Click the "Arrangement" tab
- Make sure "Mirror Displays" is unchecked

Using PC

- Press $\square$ Win + P, then select "Extend"
- Or: Right-click on the Desktop > Display Settings
- Under Multiple Displays, select "Extend these displays"

Connecting Your Laptop to Present



KubeCon



CloudNativeCon

Europe 2026



Speaker Tech Check - Ready Room Booking



Wednesday, March 25th, 2026

10:30 AM – 10:45 AM

Europe / Amsterdam

15 minutes

E101

Dear Satyam,

You'll receive a confirmation email with the booking details shortly.

Google Slides

- Use Google Chrome as your browser
- Click the dropdown next to 'Slideshow' > select Presentation Display Options
- Check "Presenter View" and "Full Screen"
- If prompted, select "Allow"
- Choose the external display (may have a strange name like PanasonicXYZ)
- Click Start Slideshow
- <https://www.youtube.com/watch?v=-GT7WCvPcys>



KubeCon



CloudNativeCon

Europe 2026

Fusing FinOps, Forecasting, and Kubernetes at Scale

Ankur Singh - Red Hat
Satyam Bhardwaj - Mirantis



WHO ARE WE?

Ankur Singh

Senior Site Reliability Engineer, ARO



Satyam Bhardwaj

Software Engineer, OSPO



Agenda



KubeCon



CloudNativeCon

Europe 2026

- Reactive FinOps in a dynamic world
- The Pattern - Observe - Forecast - Recommend
- The Architecture
- Extensibility & Scalability
- Demo



KubeCon



CloudNativeCon

Europe 2026

**The best time to forecast your
cloud spend was six months ago.
The second best time is now!**



FinOps Today: Looking at Last Month's Receipts



KubeCon



CloudNativeCon

Europe 2026

- Bill arrives → find the waste → file a ticket → wait → repeat
- **8-week feedback loop** in an industry that deploys hourly
- Autoscalers react to what's happening *right now* — nobody plans for *next week*

56.5%

still set limits by hand

State of FinOps 2026



8-week feedback loop

Plenty of Pieces, No Glue



KubeCon



CloudNativeCon

Europe 2026

Where Metrics Live

Prometheus

OpenTelemetry

OpenCost

Thanos / Cortex

Where Predictions Happen

Toto

ARIMA

TimesFM

Where Decisions Are Made

Karpenter

OpenCruise

KEDA

Cast AI



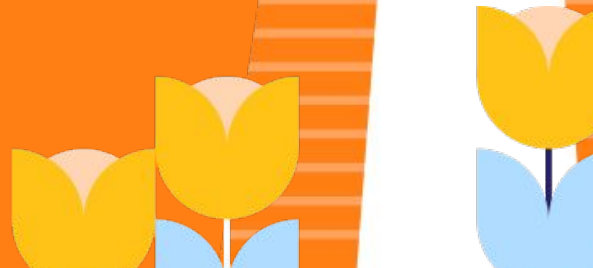
KubeCon



CloudNativeCon

Europe 2026

**What if your cluster could tell
you what it will need - before
the bill arrives?**



From Dashboards to Agents



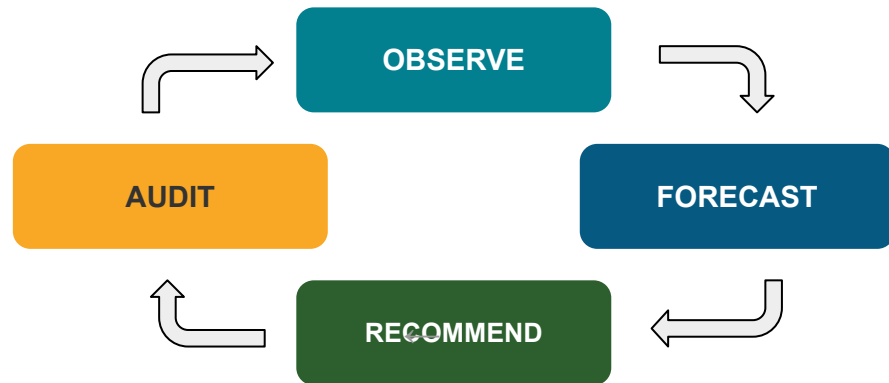
KubeCon



CloudNativeCon

Europe 2026

- **OBSERVE** - Ingest normalized cluster telemetry from diverse sources
- **FORECAST** - Multi-horizon predictions with uncertainty bounds (P10/P50/P90)
- **RECOMMEND** - Actionable policies: autoscaler configs, capacity plans, node removal
- **AUDIT** - Full traceability, human-in-the-loop approval, GitOps-native



A Kubernetes operator - a reconciliation loop that runs continuously.



KubeCon



CloudNativeCon

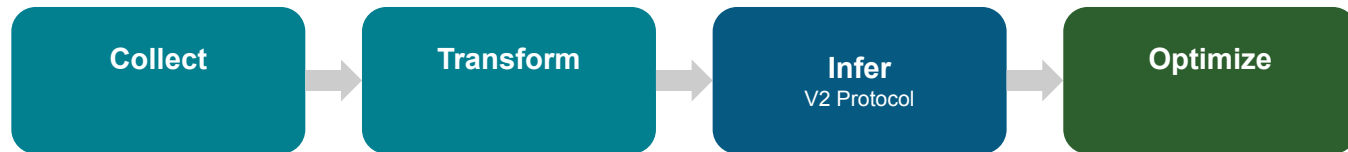
Europe 2026

Three CRDs.
Four Pipeline Stages.
One **kubectl apply**.



ForecastPipeline CR

ForecastModel CR



Output: ForecastRecommendation CRs

Grafana Dashboard

- **ForecastPipeline** => *what* to forecast
- **ForecastModel** => *how* to forecast
- **ForecastRecommendation** => *what to do about it*



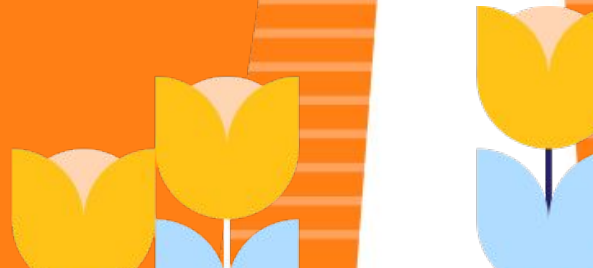
KubeCon



CloudNativeCon

Europe 2026

**Garbage In, Garbage Out.
So let's talk about how the data
flows clean**



- Standard data formats
 - Sources output MetricFrame. Predictions output ForecastEnvelope
 - Any source works with any model - no custom glue
- Multi-model Inference
 - Same pipeline, swap the model - Toto, ARIMA, TimesFM, Prophet
 - Compare models on *your* data, pick the best one
- Self-checking accuracy
 - Compares past predictions vs what actually happened (hindcasting)
 - System tells you when to stop trusting it
- Human-in-the-loop
 - Recommendations need human approval, fir into GitOps workflows
 - Build trust first, automate later



KubeCon



CloudNativeCon

Europe 2026

**The best part? You can bring
your own everything.**



Plugin Architecture + Multi-Cluster

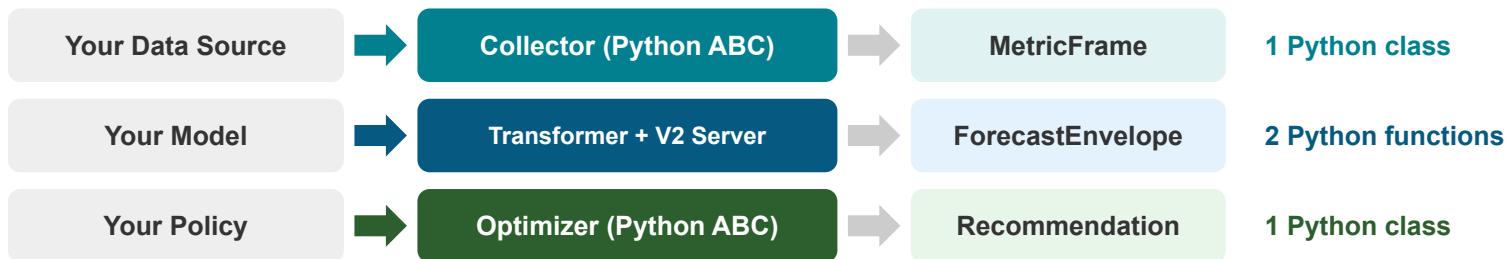


KubeCon



CloudNativeCon

Europe 2026



- **Scale across clusters:** Runs on one cluster, queries many via Prometheus federation (Thanos/Cortex/Promxy)
- **Right language for the job:** Go handles K8s plumbing. Python handles data and ML. Contributors only need Python.
- **No core changes needed:** Your plugin works with every other plugin automatically via Python entry_points



KubeCon



CloudNativeCon

Europe 2026

Let's see it run!



What's Next ?

Now (v0.1)	Next	Future
Single Cluster	Multi-cluster	More Model support
Toto + ARIMA	Hindcasting loop	HPA tuning optimizer
Idle node removal recommendation	External Database	Workload-level forecasting
Grafana dashboards + Forecast API	Plugin registry	Helm chart & OLM catalog

github.com/augur-k8s/augur | **Apache-2.0**



KubeCon



CloudNativeCon

Europe 2026



Thank You

