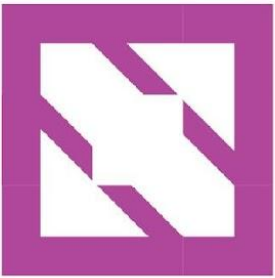




**KubeCon**



**CloudNativeCon**

**Europe 2025**



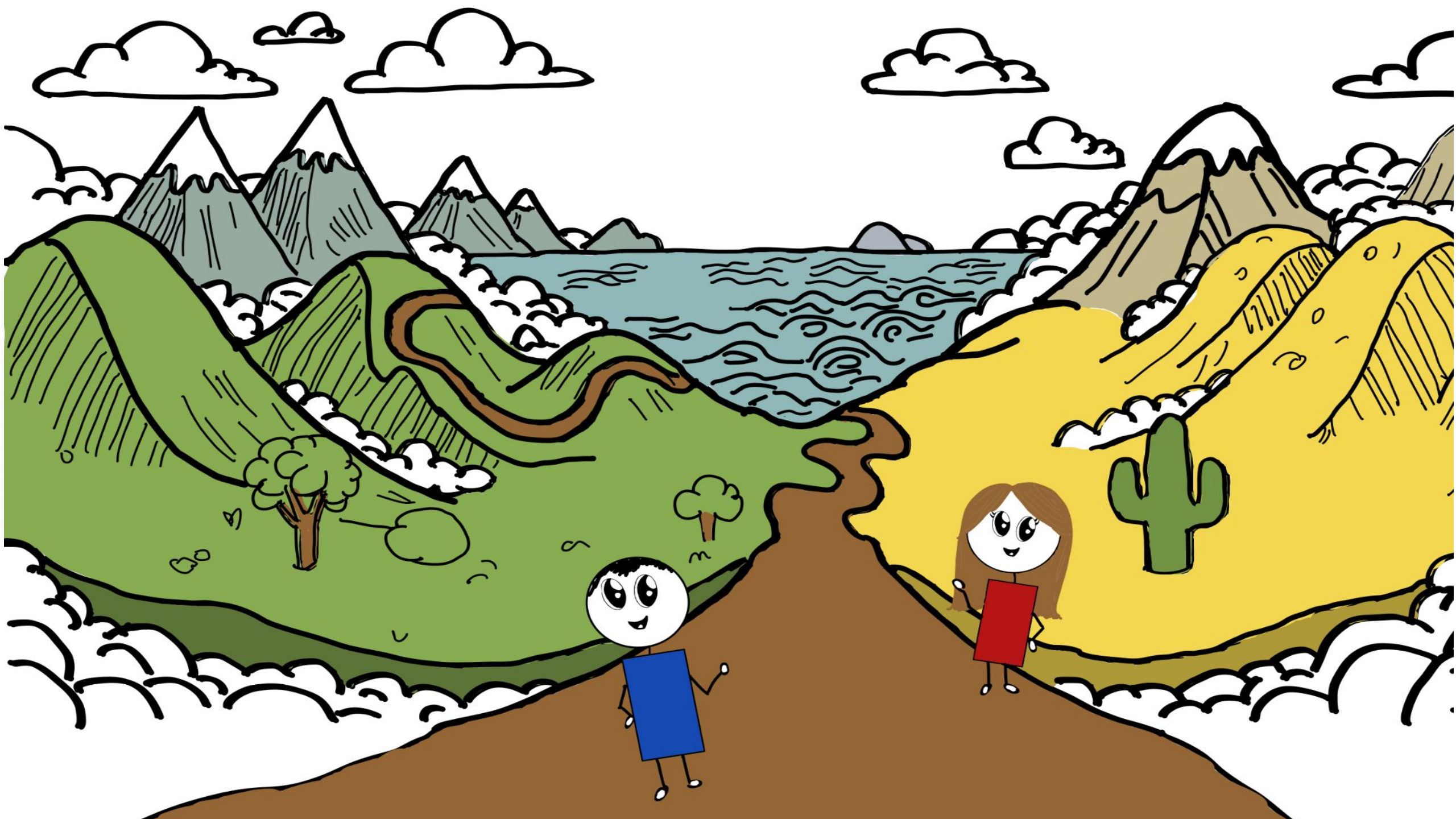


**Alexa Griffith**  
Senior Software Engineer  
Bloomberg



**Max Körbacher**  
Cloud Native Advisor & Managing Director  
Liquid Reply





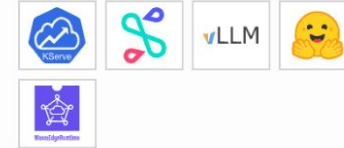
# AI Happens alongside the Cloud Native world

CNAI

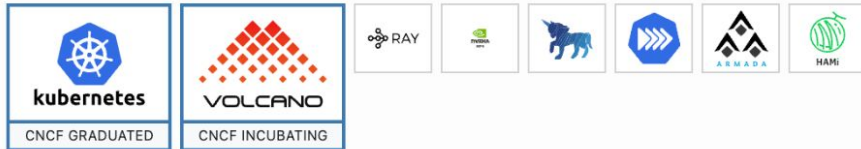
## Data Architecture



## ML Serving



## General Orchestration



## CI/CD - Delivery



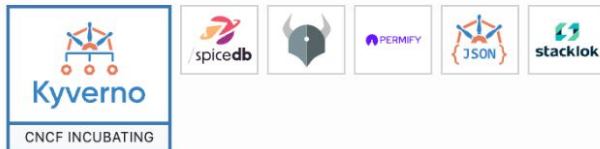
## Distributed Training



## Workload Observability



## Governance, Policy & Security



## AutoML



## Data Science



## Vector Databases



## Model/LLM Observability

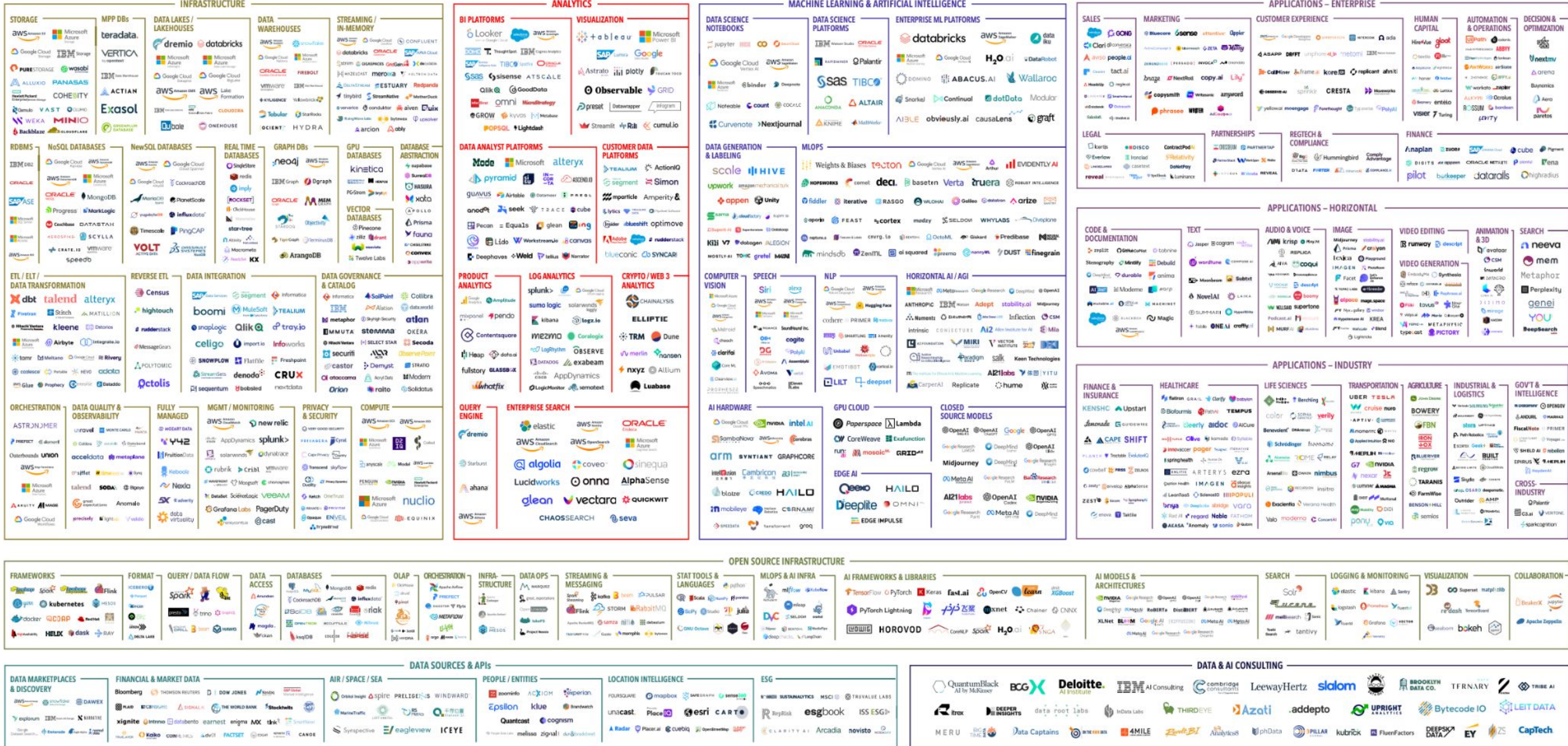


## Open Enterprise AI Blueprints



# The 2023 MAD Landscape

## THE 2023 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE





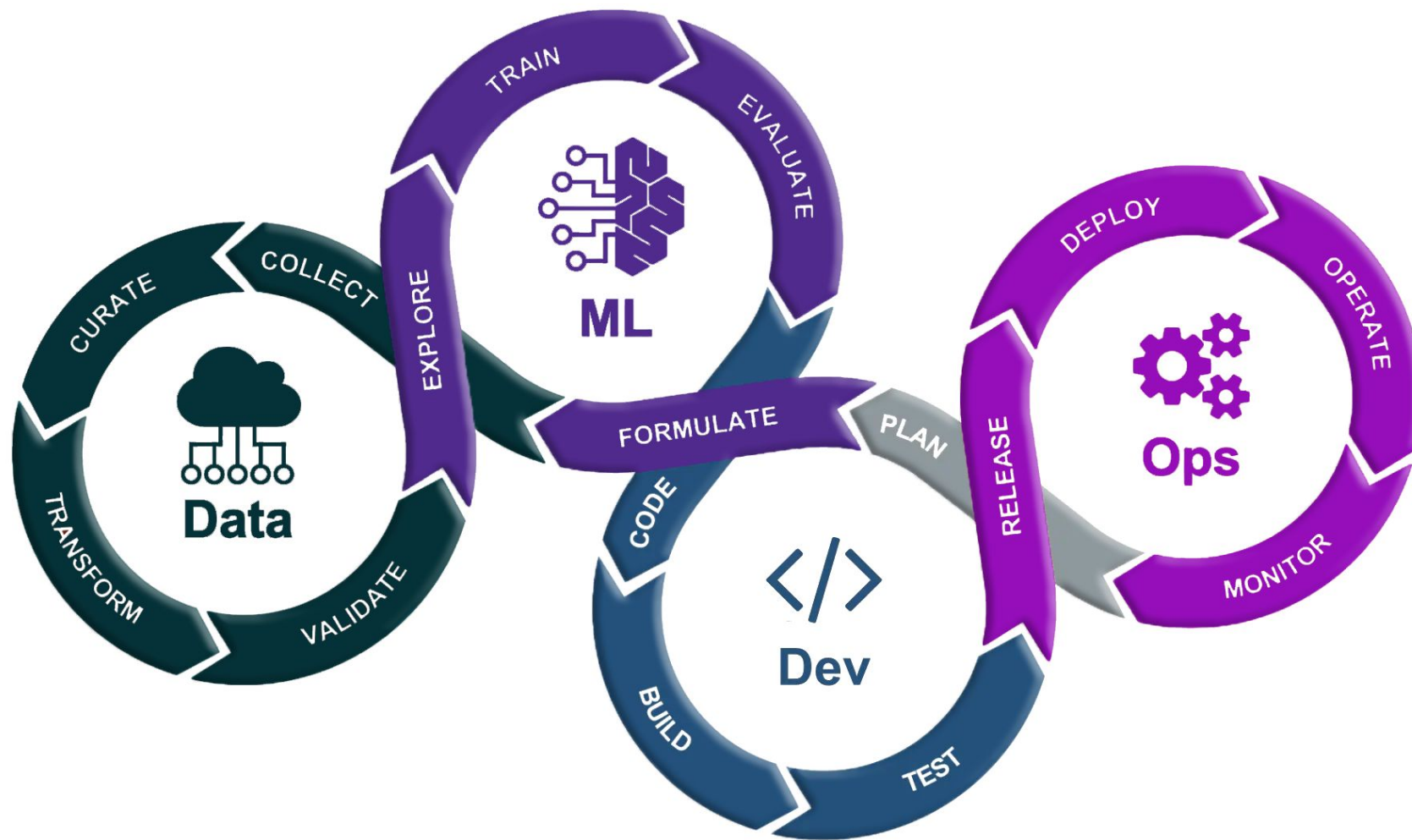
# Kubernetes is a platform to build platforms

2020 - Infrastructure Abstraction & Automation

2022 - Going to Edge, Telco & Fleets

2023 - Internal Development Platforms (IDP)

**AI/ML/LLM in Cloud Native**





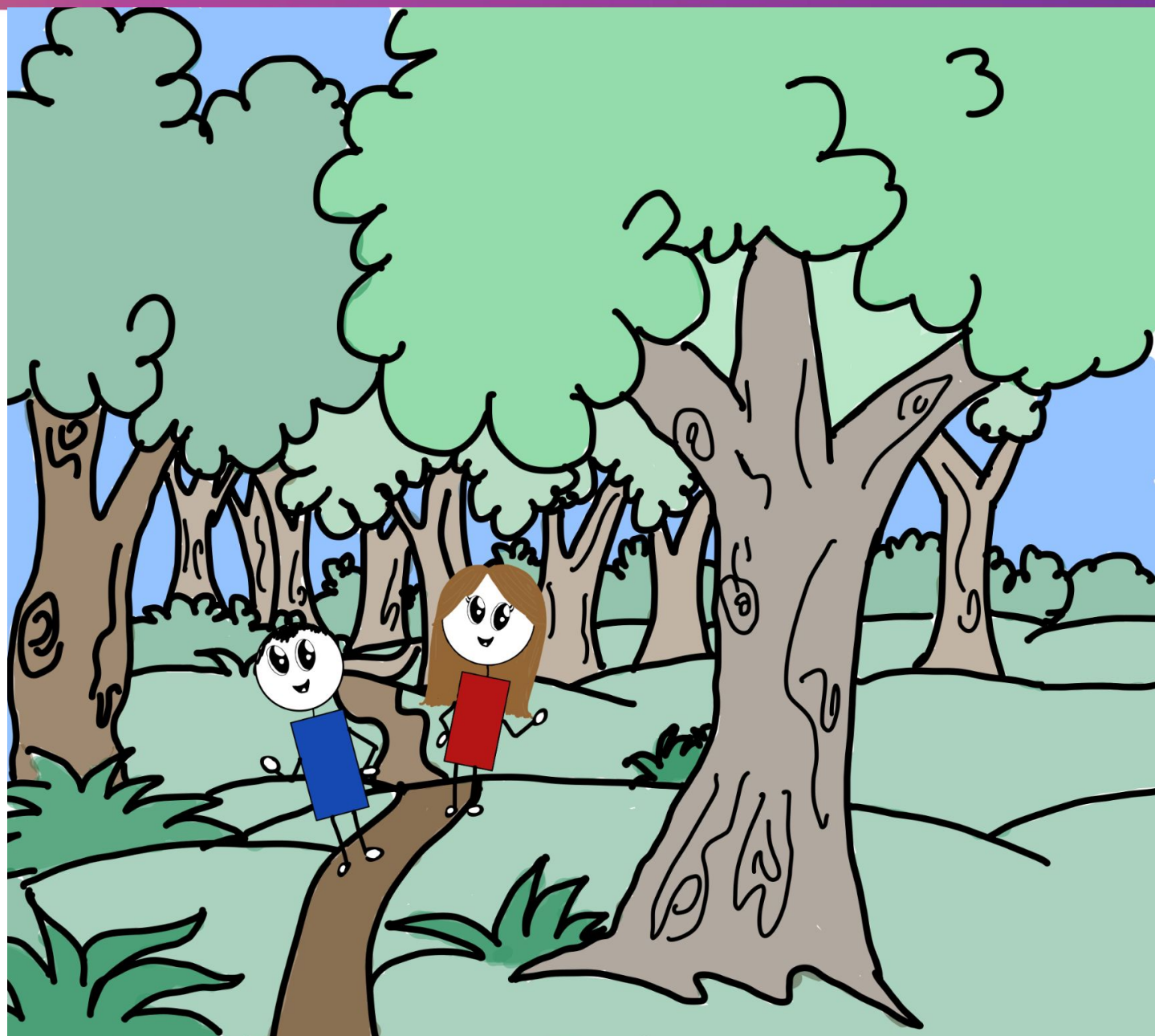
KubeCon



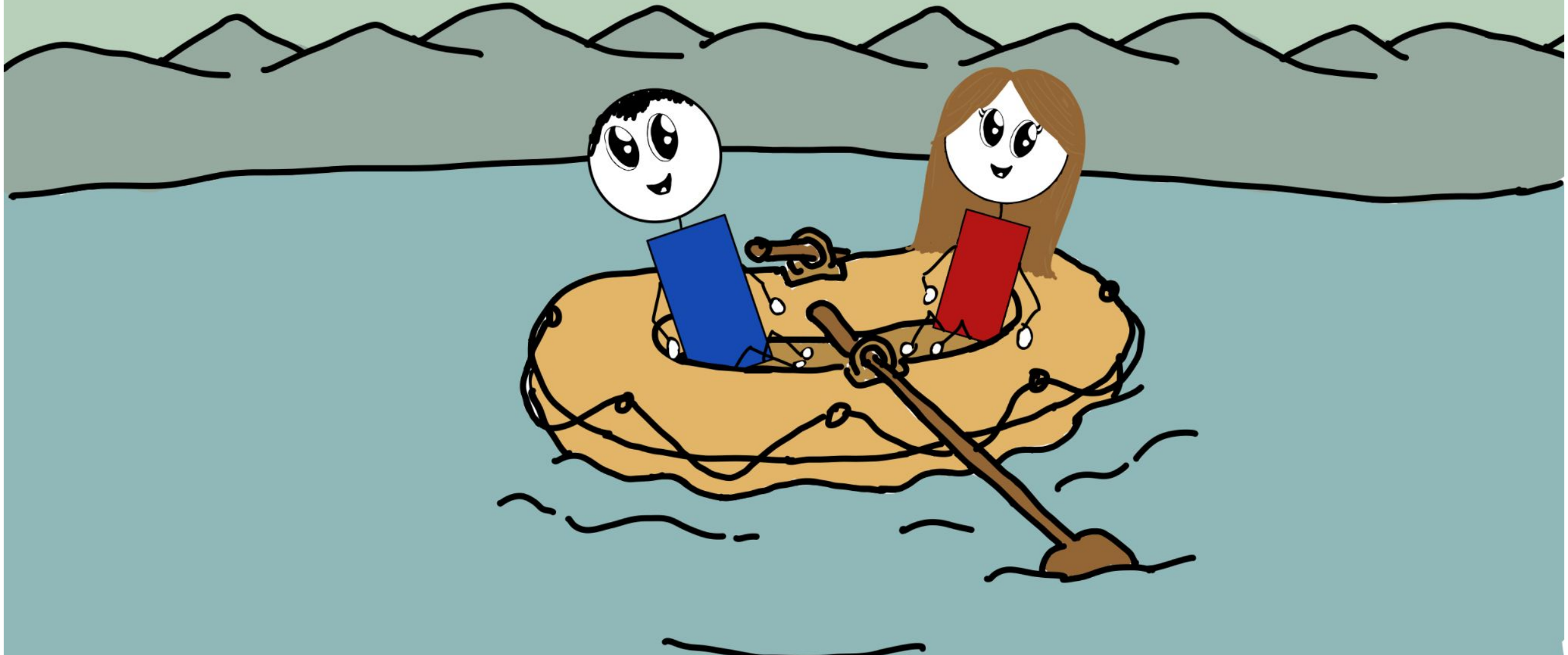
CloudNativeCon

Europe 2025

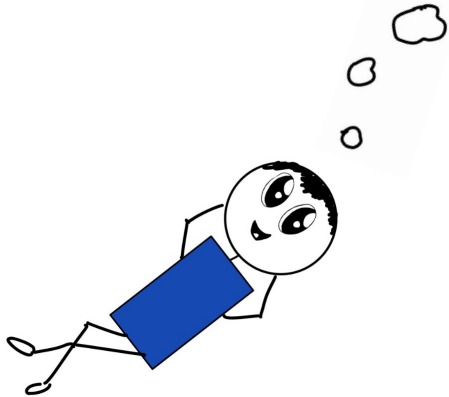
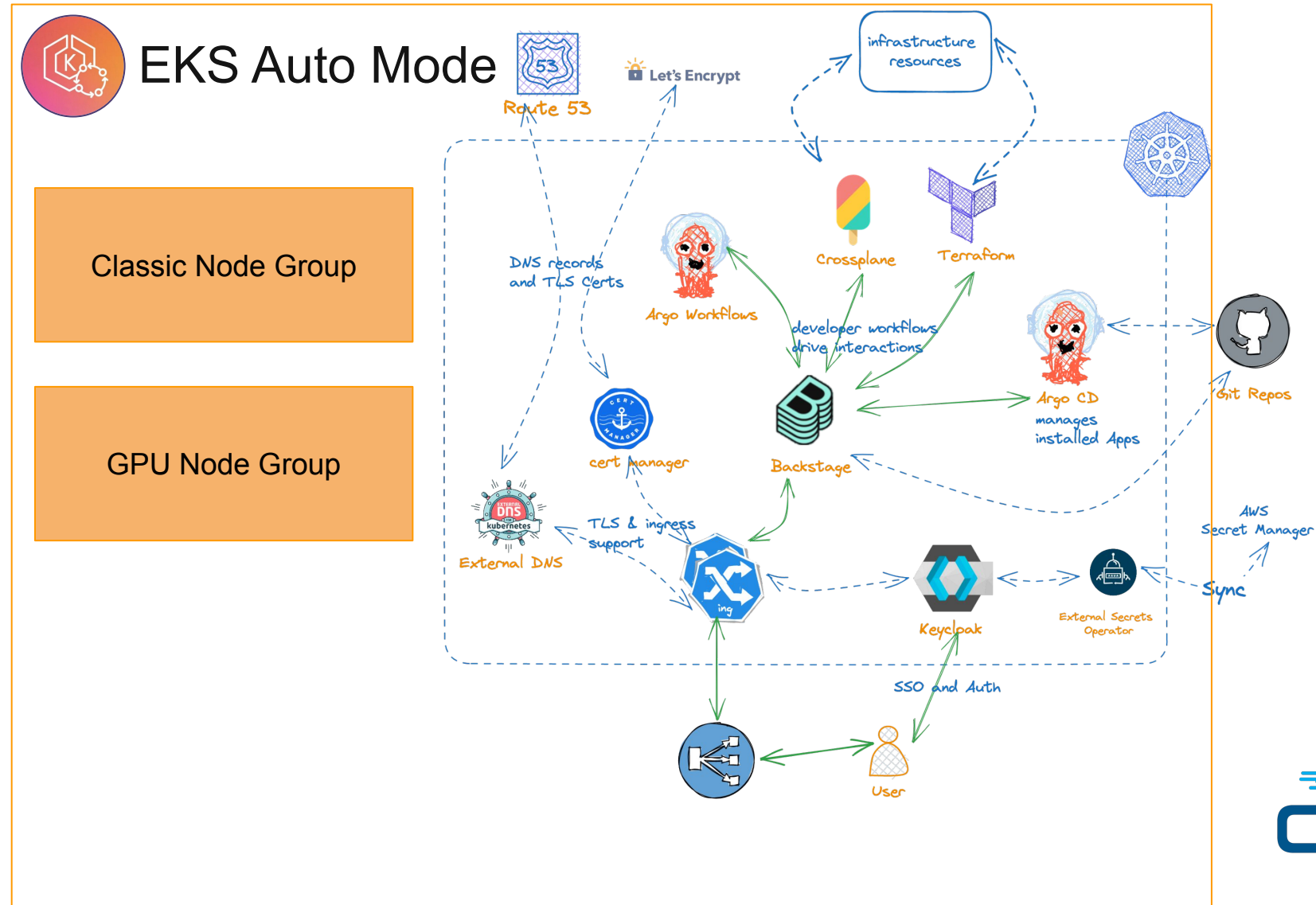
What are my options  
to run GenAI  
workloads and  
enable others to do  
so too?



# Stage 1: TVP



# “Minimal IDP”



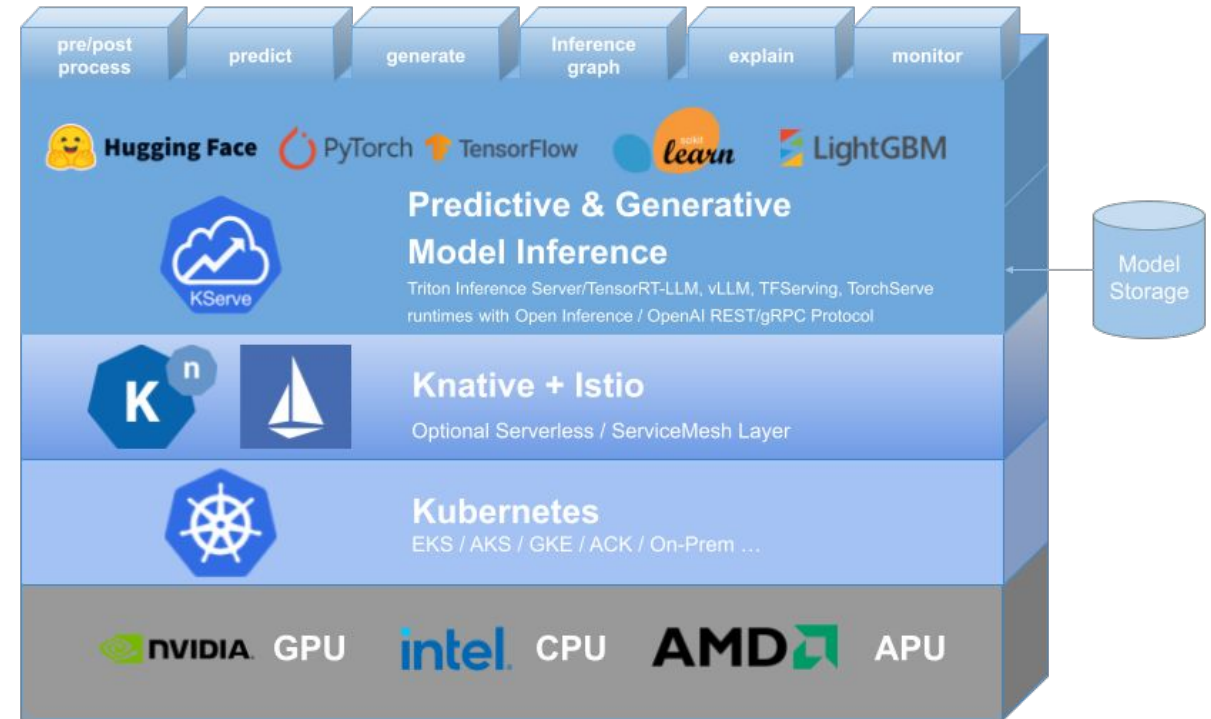


# Stage 2: MVP



# KServe: Powering the AI Expedition

- Simplifies scalable model serving on Kubernetes
- Supports multiple model runtimes out of the box, including LLMs

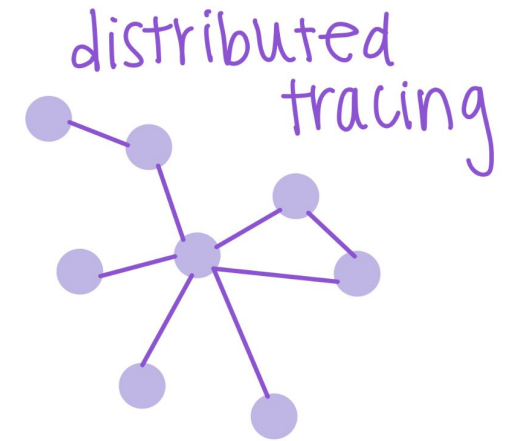
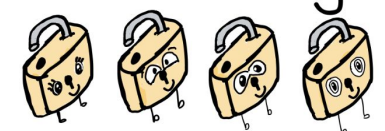


# KServe: Powering the AI Expedition

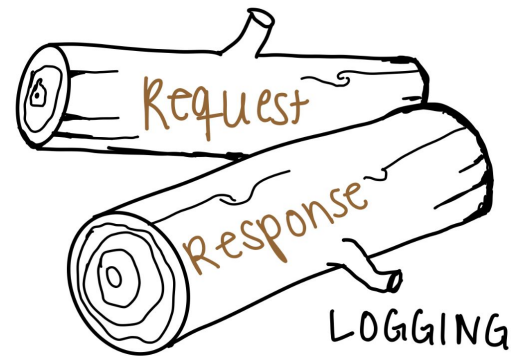
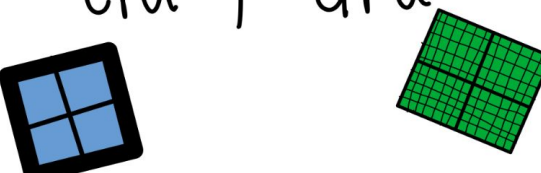
scale  
↓ to & ↑ from  
zero



security  
with  
AuthN/AuthZ



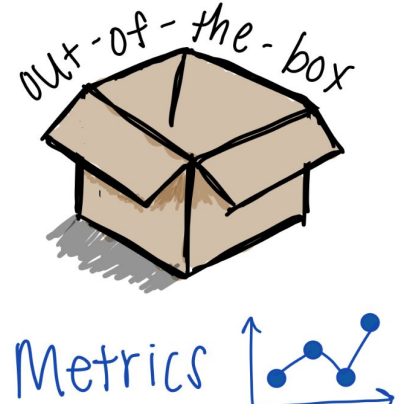
request based  
autoscaling  
on  
CPU | GPU



traffic management

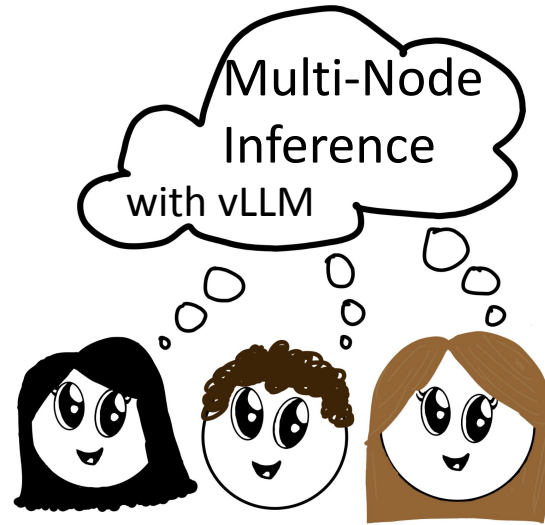


out-of-the-box  
Metrics

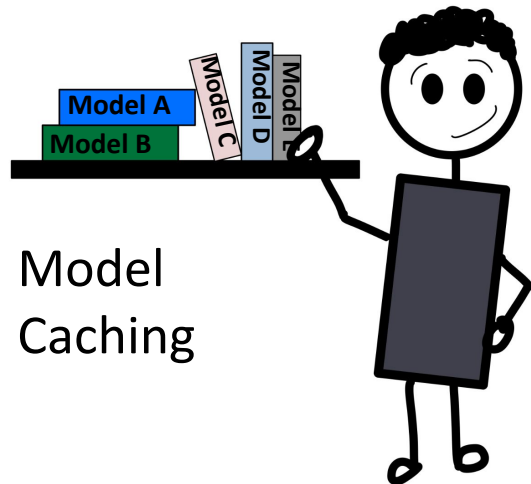
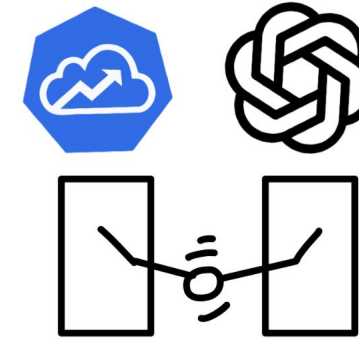


# Building the AI Vessel: KServe for GenAI

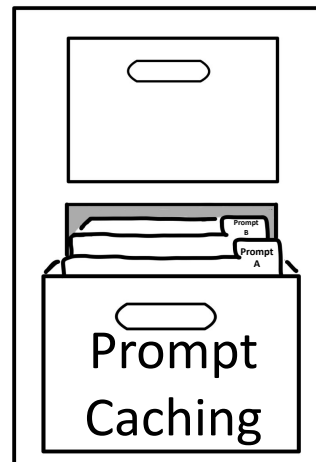
LLM  
Metric-based  
Autoscaling



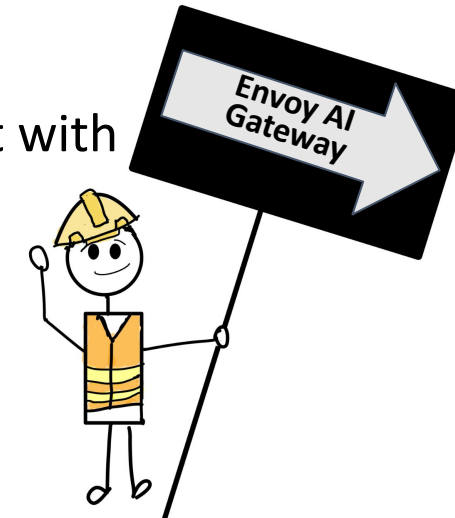
OpenAI Protocol Support



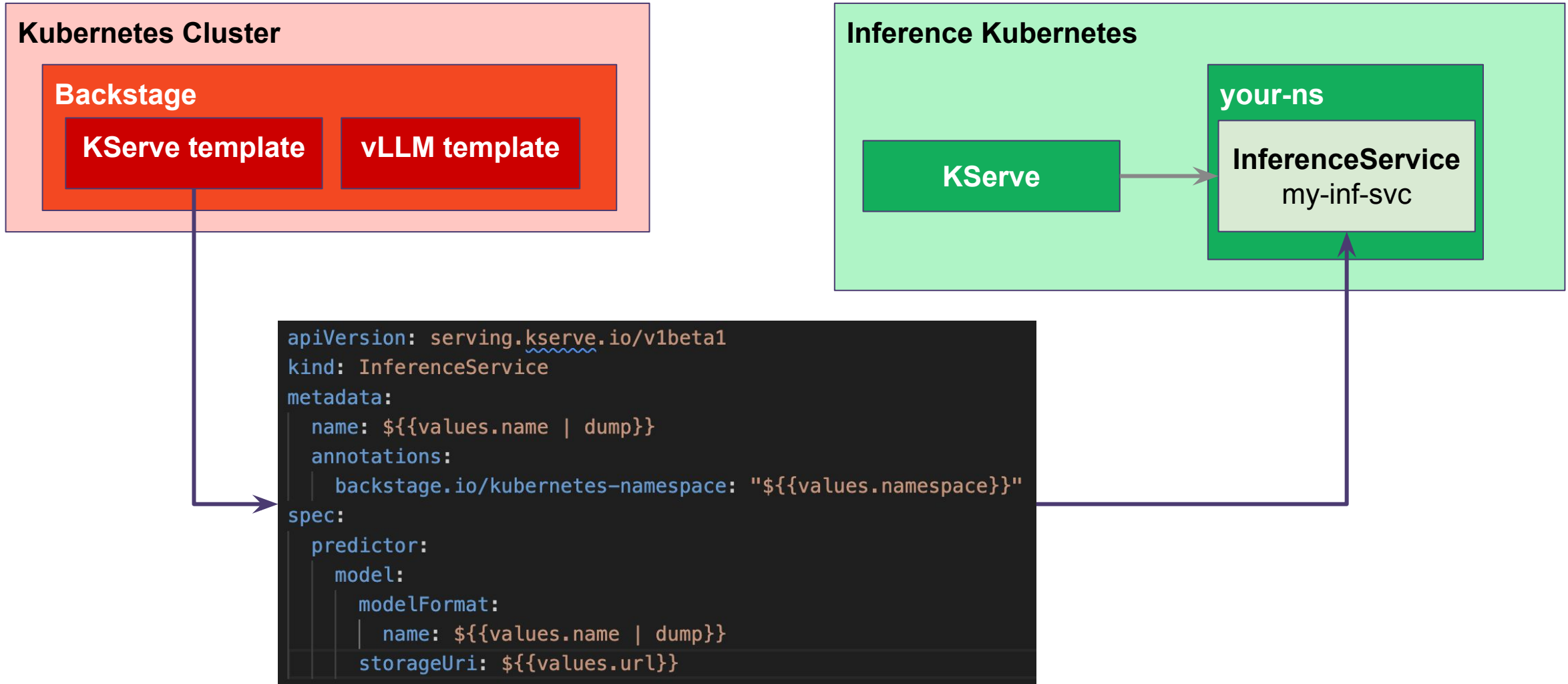
Model  
Caching



Traffic  
Management with



# Utilize Backstage for Inferencing & More



# Navigating Access with Envoy AI Gateway

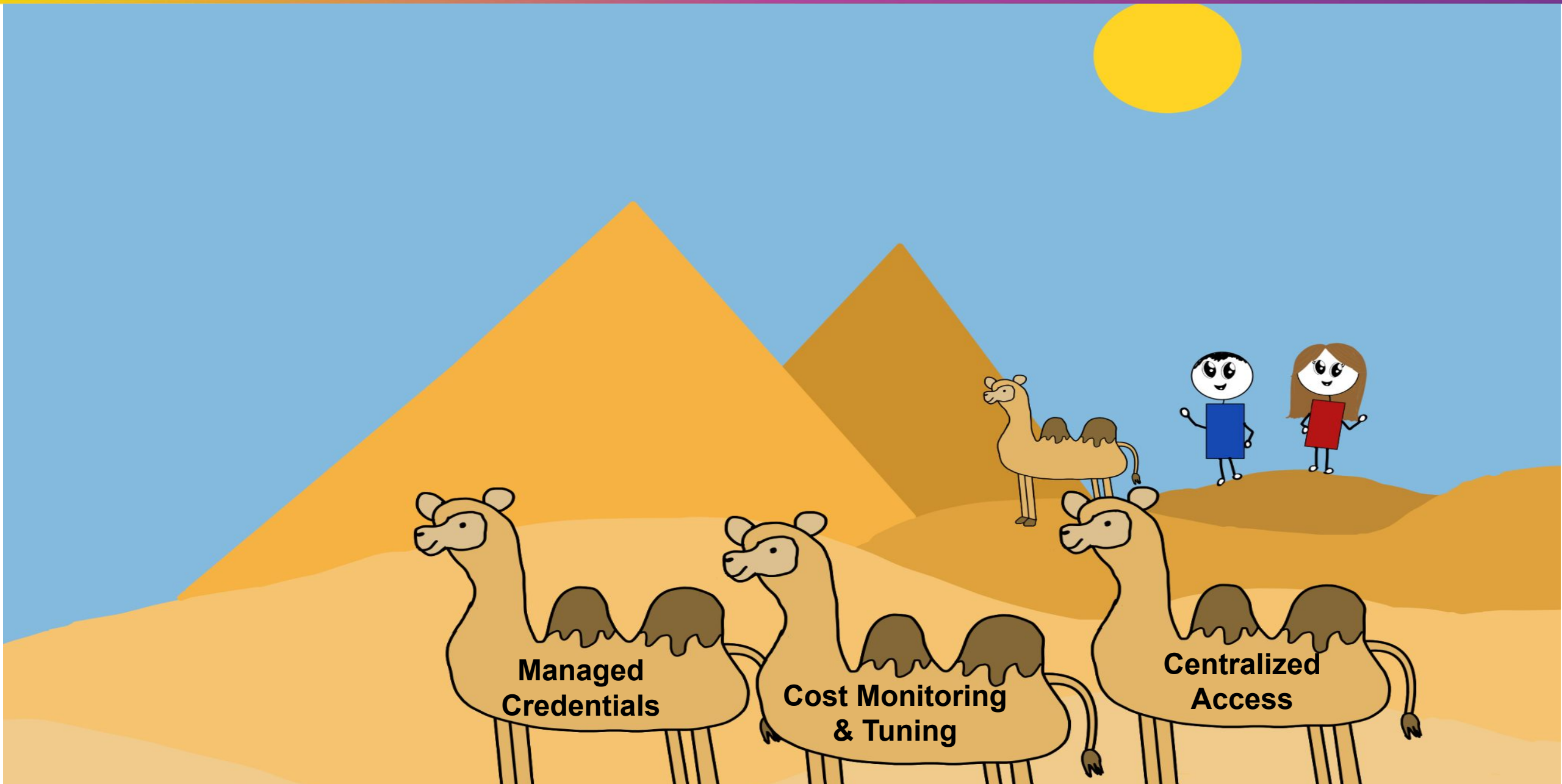


KubeCon



CloudNativeCon

Europe 2025

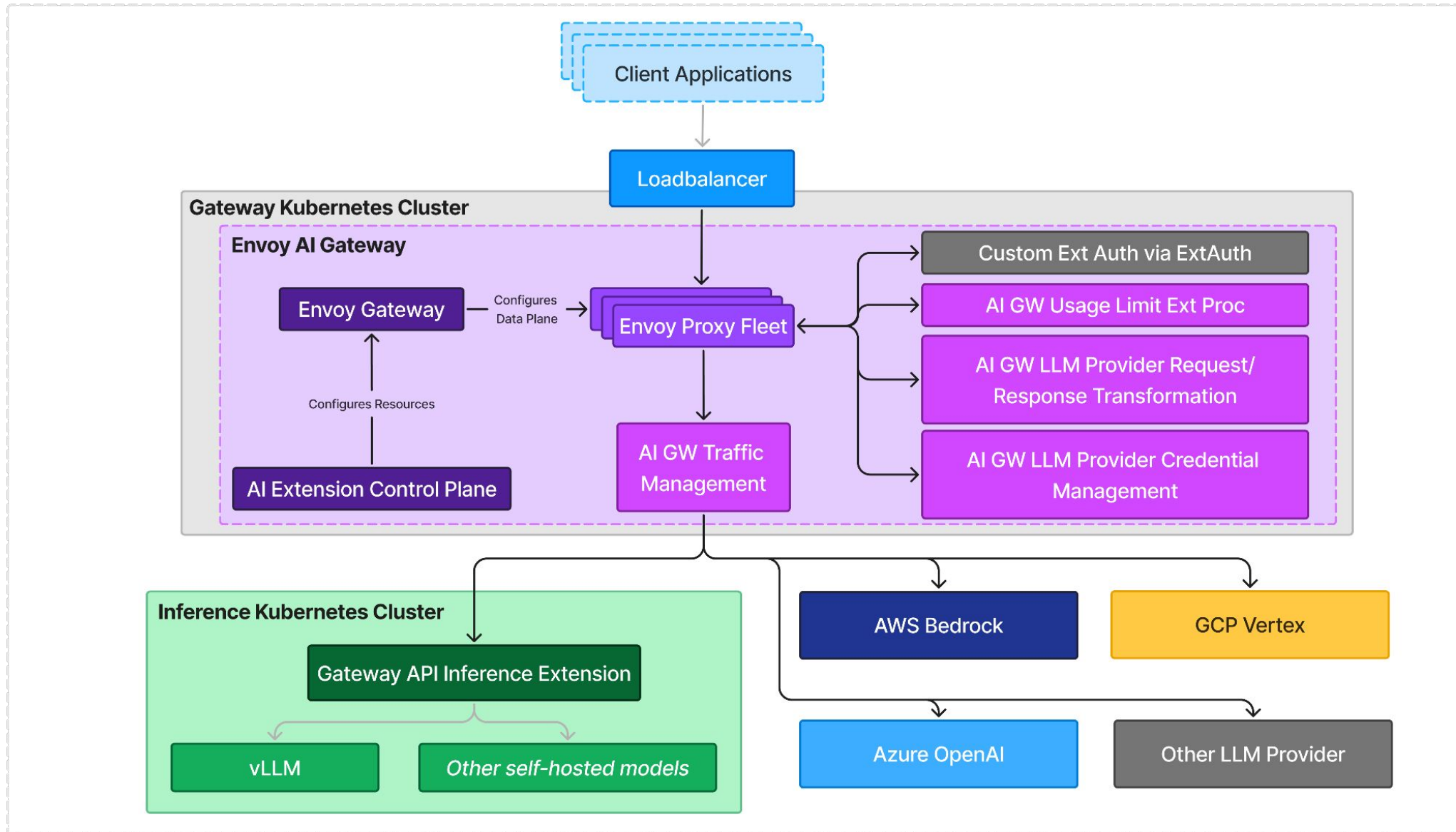


**Managed  
Credentials**

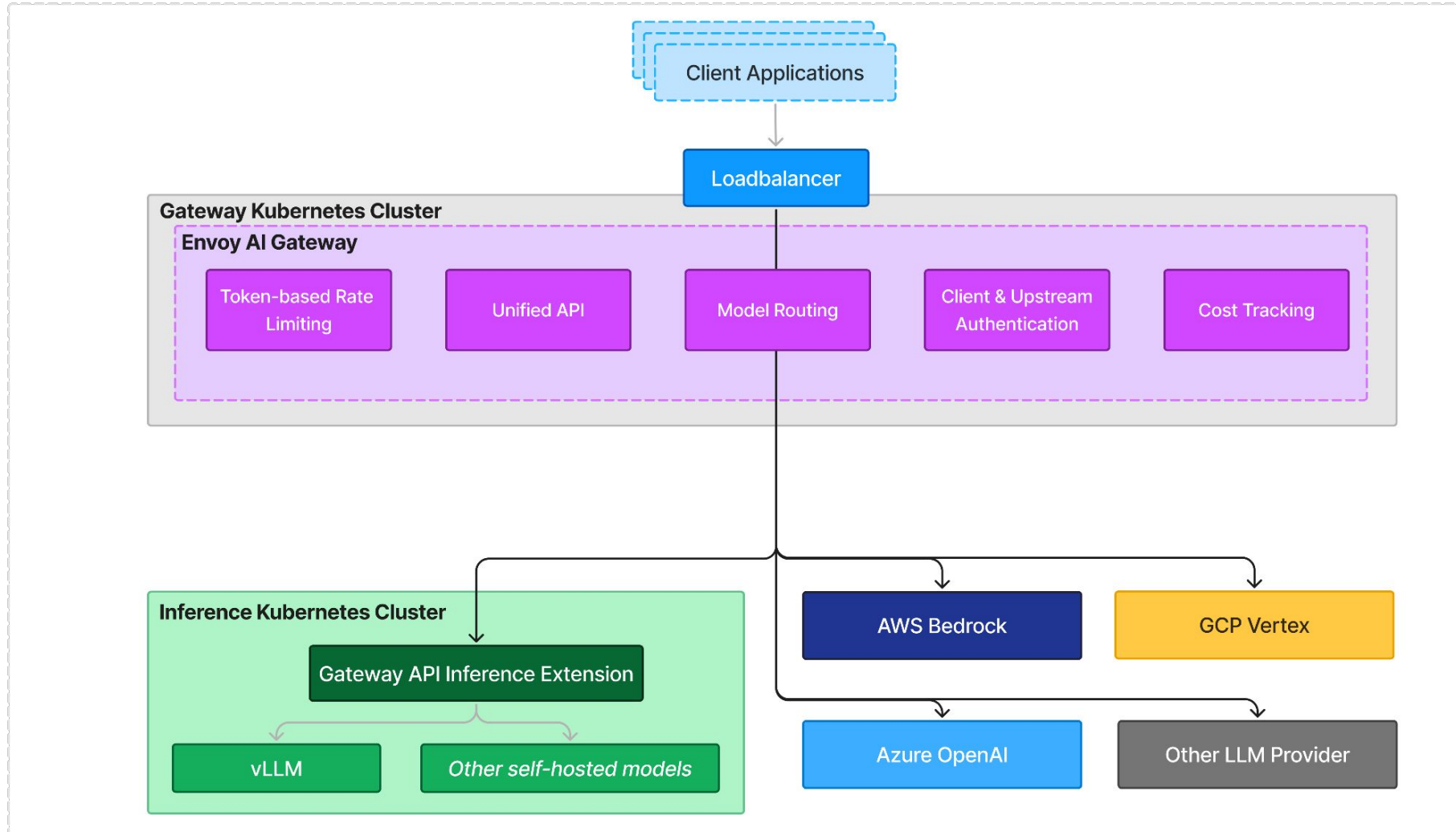
**Cost Monitoring  
& Tuning**

**Centralized  
Access**

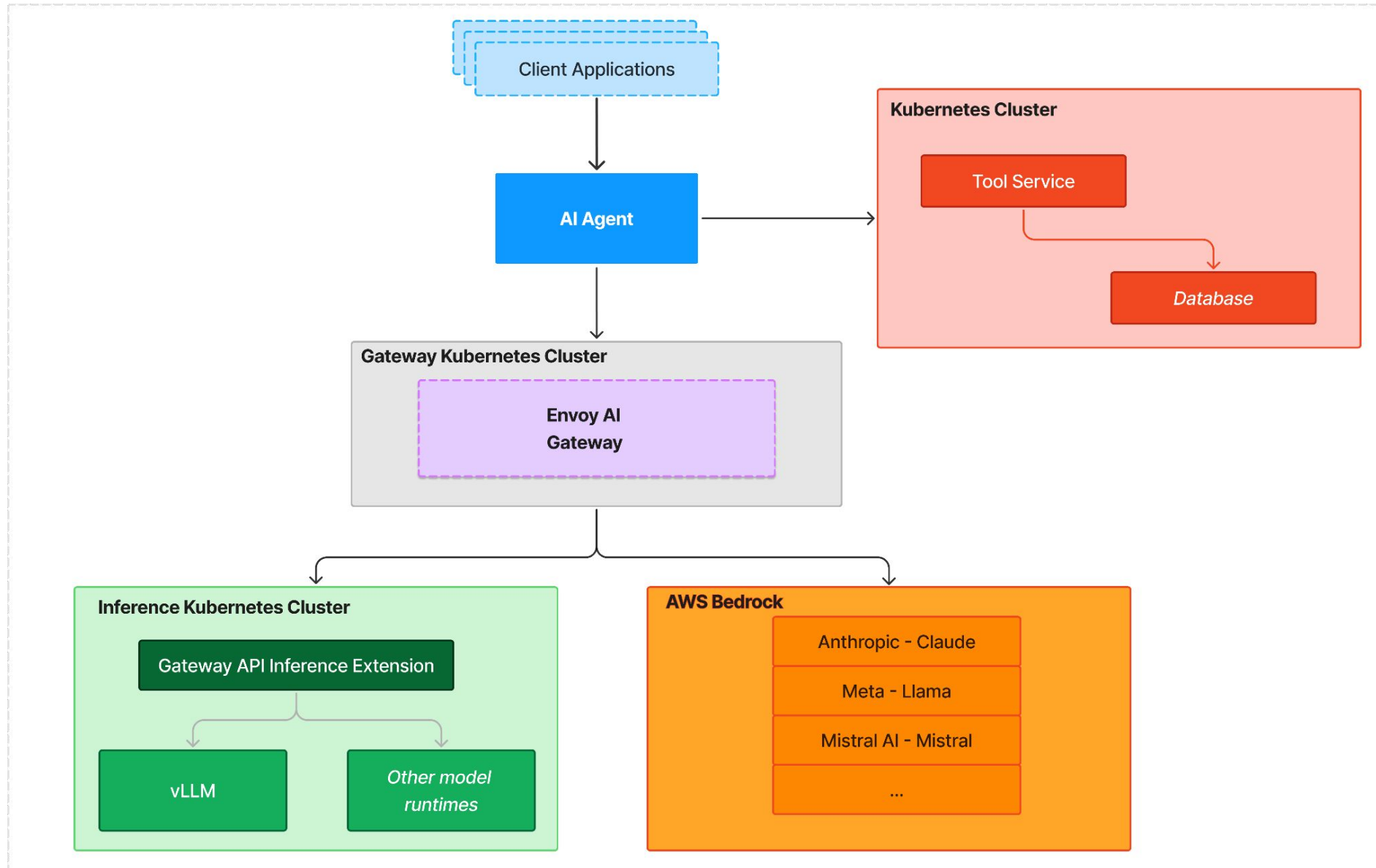
# Through the Envoy AI Gateway



# Through the Envoy AI Gateway



# Map of the Route Ahead



# AI Gateway Demo



KubeCon



CloudNativeCon

Europe 2025

```
EXPLORER
PLATFORM-DEMO-KUBE...
go.mod
go.sum
main.go M
README.md

main.go M X
~/github/platform-demo-kubecon-eu-2025/main.go • Modified
1 package main
2 import (
3     "context"
4     "flag"
5     "log"
6     "encoding/json"
7
8     openai "github.com/openai/openai-go"
9     "github.com/openai/openai-go/option"
10 )
11 var (
12     useAIGateway = flag.Bool("use-ai-gateway", true, "Use AI Gateway instead of direct Bedrock")
13     aiGatewayURL = flag.String("ai-gateway-url", "http://localhost:8080", "AI Gateway URL")
14     awsSessionToken = flag.String("aws-session-token", "", "AWS Session Token (optional)")
15     modelName = flag.String("model-name", "eu.anthropic.claude-3-5-sonnet-20240620-v1:0", "Bedrock model name")
16     toolURL = flag.String("tool-url", "", "External tool URL for weather service")
17 )
18 const question = "What is the weather in New York City?"
19
20 func main() {
21     flag.Parse()
22
23     // Determine base URL (AI Gateway or Bedrock)
24     baseURL := ""
25     if *useAIGateway {
26         log.Println("Using AI Gateway for requests.")
27         baseURL = *aiGatewayURL + "/v1/"
28     } else {
29         log.Println("Using Amazon Bedrock for requests.")
30     }
31
32     // Initialize OpenAI client
33     client := openai.NewClient(
34         option.WithBaseURL(baseURL),
35     )
36
37     params := openai.ChatCompletionNewParams{
38         Messages: openai.F([]openai.ChatCompletionMessageParamUnion{
39             openai.UserMessage(question),
40         }),
41         Tools: openai.F([]openai.ChatCompletionToolParam{
42             {
43                 Type: openai.F(openai.ChatCompletionToolTypeFunction),
44                 Function: openai.F(openai.FunctionDefinitionParam{
45                     Name: openai.String("get_weather"),
46                     Description: openai.String("Get weather at the given location"),
47                     Parameters: openai.F(openai.FunctionParameters{
48                         "type": "object",
49                         "properties": map[string]interface{}{
```

# Charting the Course for Managed LLMs



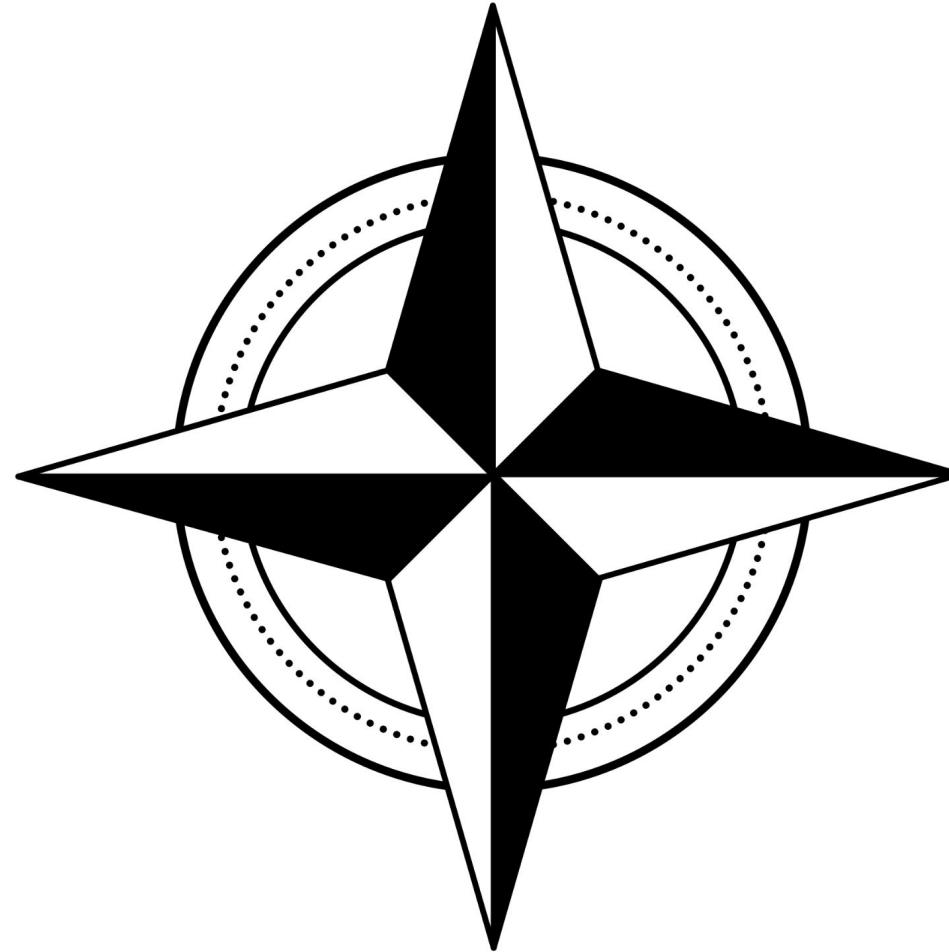
KubeCon



CloudNativeCon

Europe 2025

**AI without the Overhead**

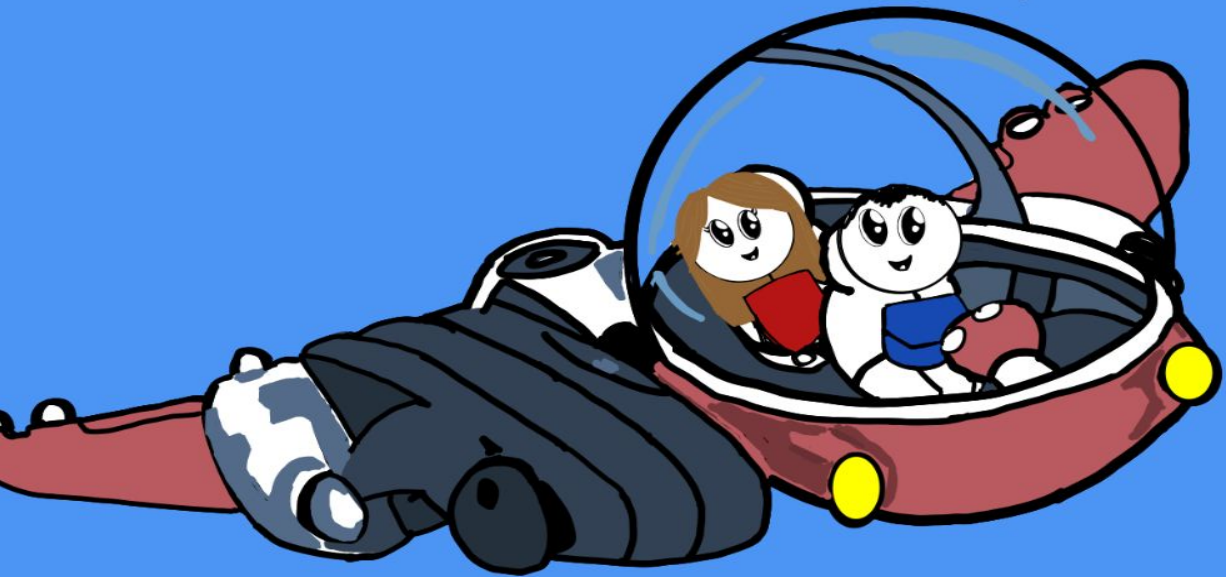


**Optimized &  
Up-to-Date Models**

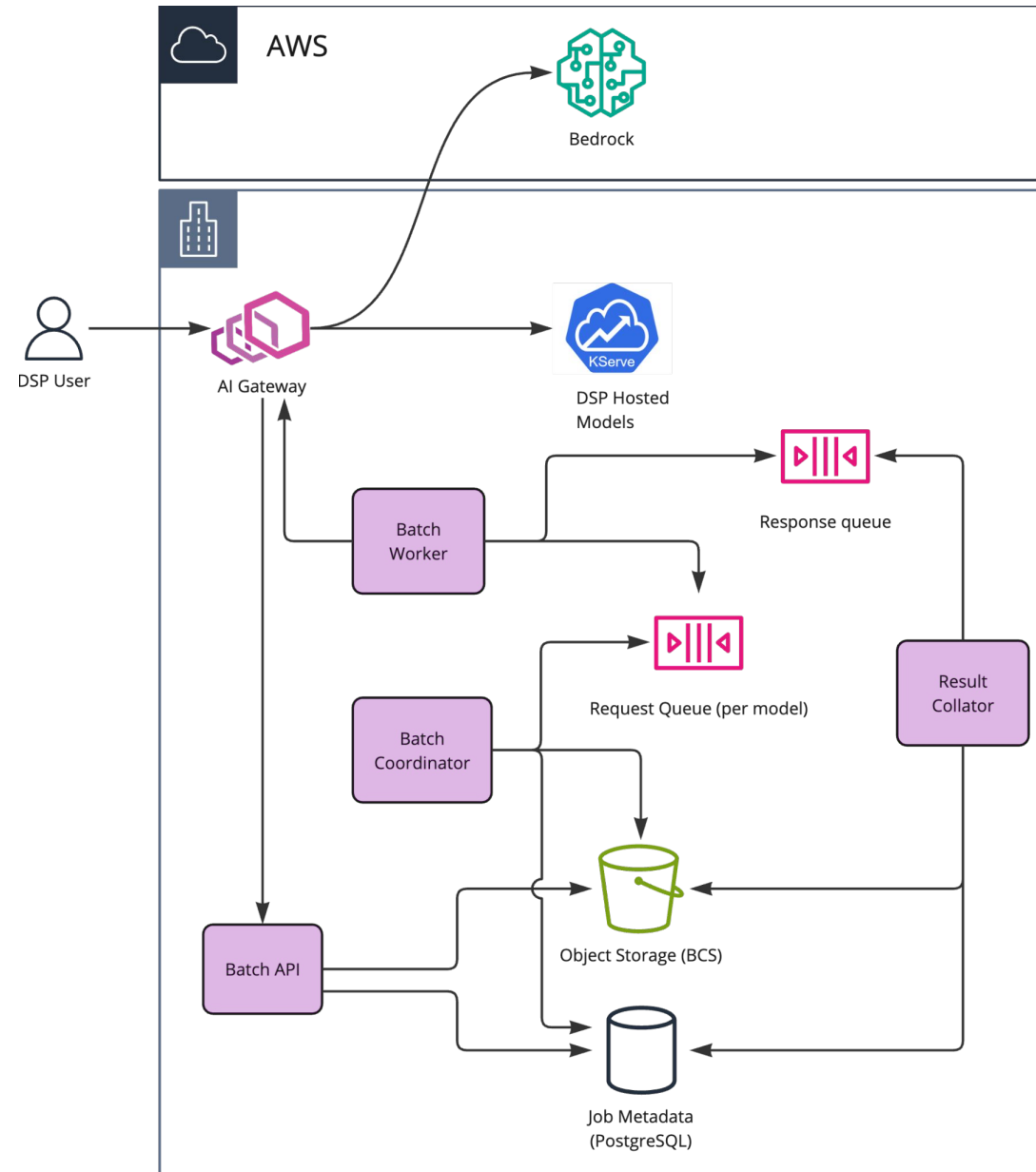
**Unified Access**

**Cloud-Native Scalability**

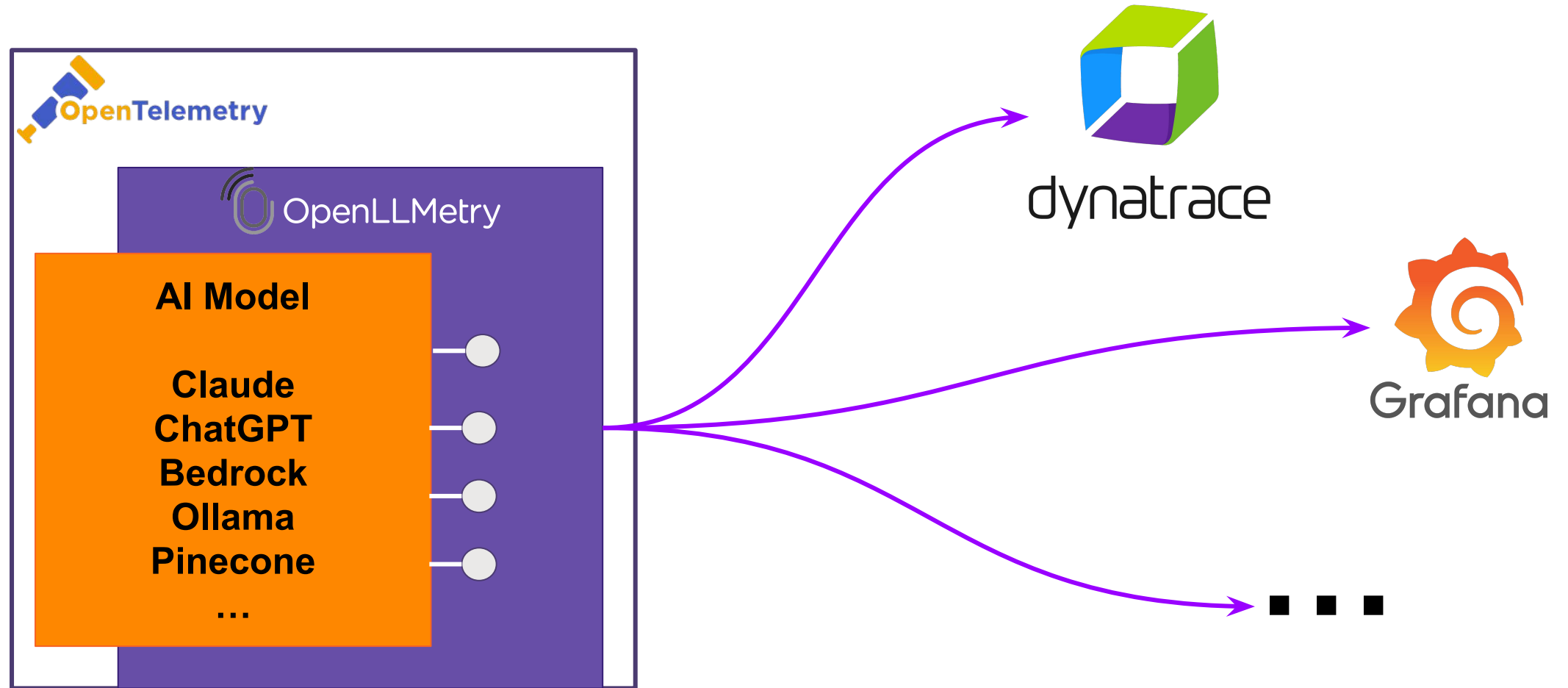
# Stage 3: Advanced Features



# Expedition Mode: Scaling AI with Batch Inference



# LLM Observability Observatory



### Service Health & Performance



Open Problems

2

Service Health



# of Total Requests

796.00  
= 150%

Cost

840.88US\$  
= 60.83%

API Request Duration

1.07s  
= 4,82%

PII Request Duration

4.49s  
= 61,05%

Request Patterns



### Service Quality & Guardrails

Guardrail Executions

7.55%

Toxicity

14.23%

PII Leaks

14.00%

Denied Topics

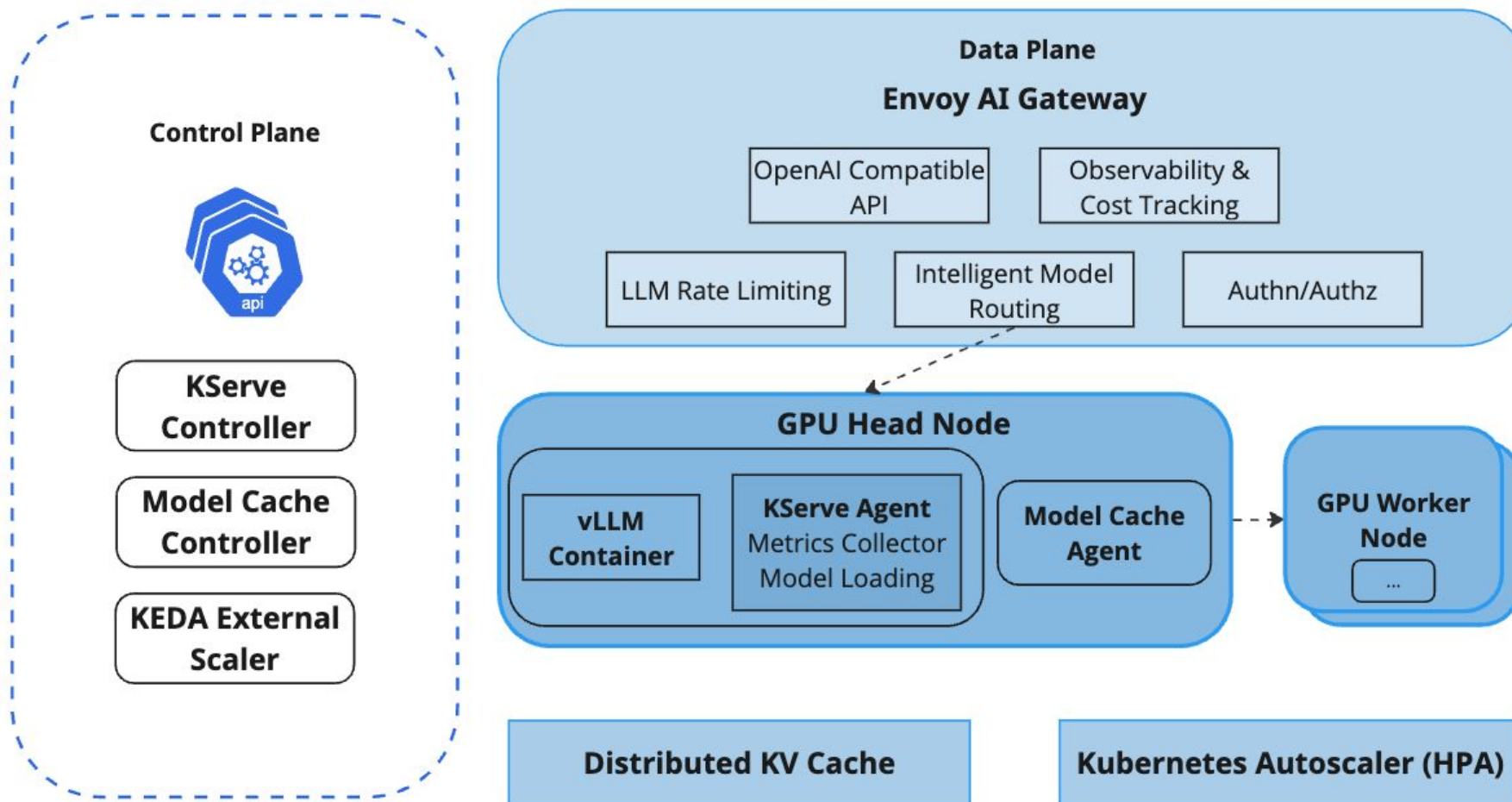
14.23%

Grounding

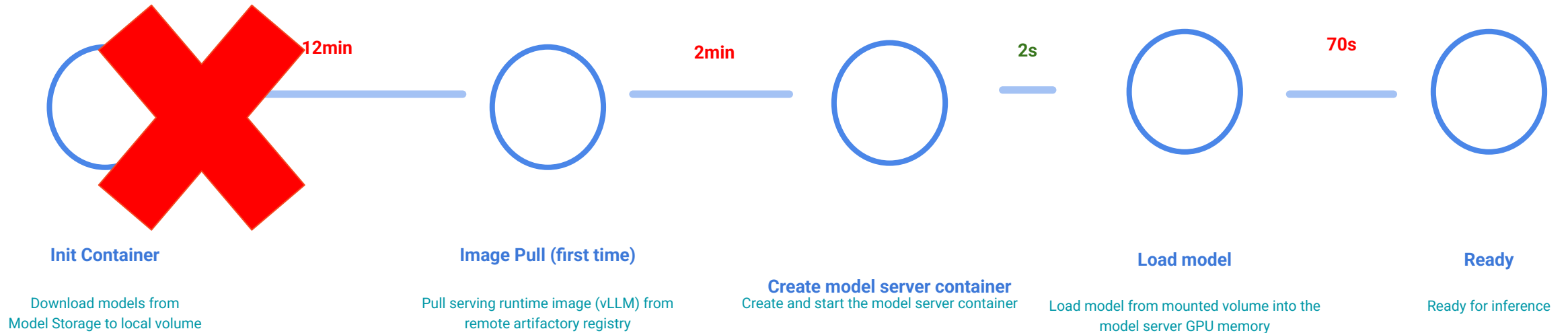
0.74  
= 3,46%



# Upgrades for the Expedition



# Model Caching



# Prompt Caching

## KVCache

Grows too fast! 

## LMCache

Reuses common prefixes.

 Faster inference!

### KV Cache Size Calculator

Select LLM Model:

deepseek-ai/DeepSeek-V3

Select data type:

float16

Enter Number of Tokens:

10000

Calculate KV Cache Size

**KV Cache Size: 16.2888 GB**

Calculation Details:

Selected Model: deepseek-ai/DeepSeek-V3  
Hidden Size: 7168  
Number of Attention Heads: 128  
Number of Hidden Layers: 61  
Number of Key-Value Heads: 128  
Head Size: 56 (Hidden Size / Attention Heads)  
Data Type Size: 2 bytes  
Total Elements:  $2 \times 61 \times 10000 \times 128 \times 56 = 8744960000$   
Total Bytes:  $8744960000 \times 2 = 17489920000$  bytes  
KV Cache Size:  $17489920000 / (1024^3) \approx 16.2888$  GB

### KV Cache Size Calculator

Select LLM Model:

deepseek-ai/DeepSeek-V3

Select data type:

float16

Enter Number of Tokens:

1000000

Calculate KV Cache Size

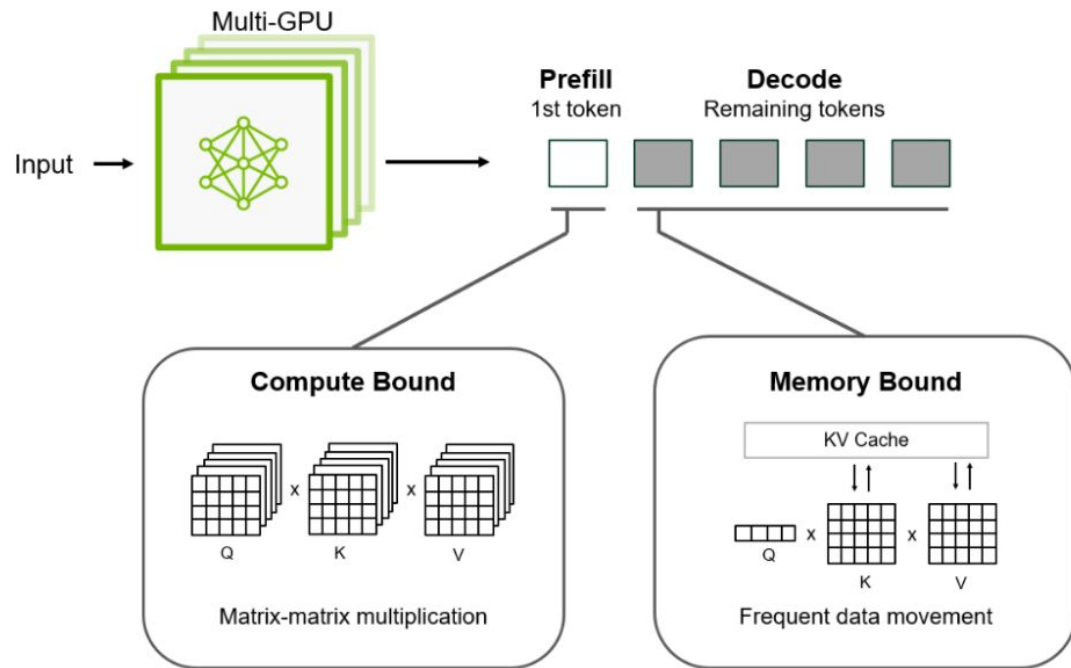
**KV Cache Size: 1628.8757 GB**

Calculation Details:

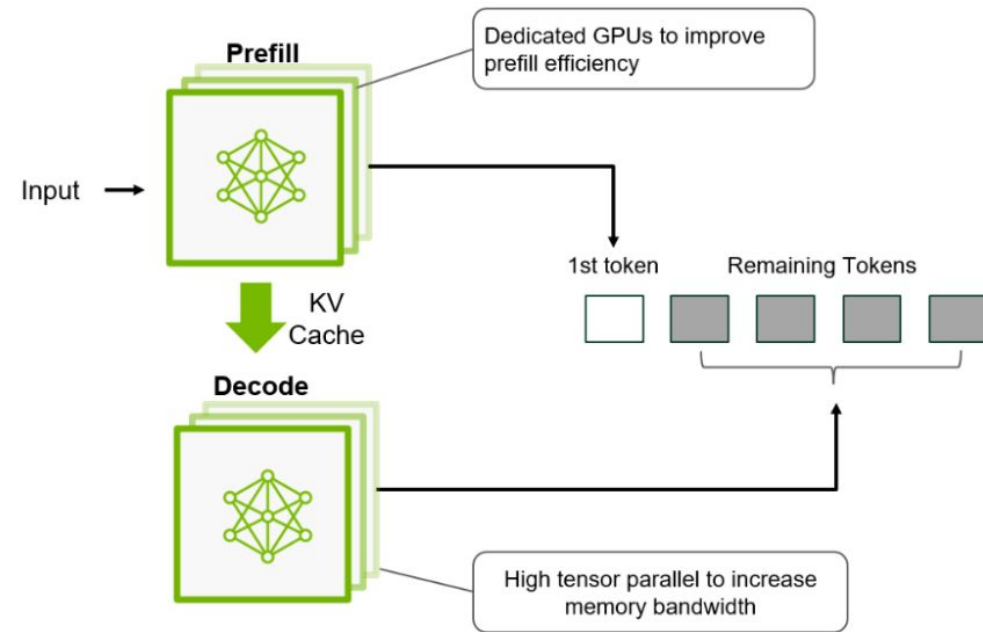
Selected Model: deepseek-ai/DeepSeek-V3  
Hidden Size: 7168  
Number of Attention Heads: 128  
Number of Hidden Layers: 61  
Number of Key-Value Heads: 128  
Head Size: 56 (Hidden Size / Attention Heads)  
Data Type Size: 2 bytes  
Total Elements:  $2 \times 61 \times 1000000 \times 128 \times 56 = 874496000000$   
Total Bytes:  $874496000000 \times 2 = 1748992000000$  bytes  
KV Cache Size:  $1748992000000 / (1024^3) \approx 1628.8757$  GB

# Disaggregated Serving

## Traditional Serving



## Disaggregated Serving



# End of the Journey

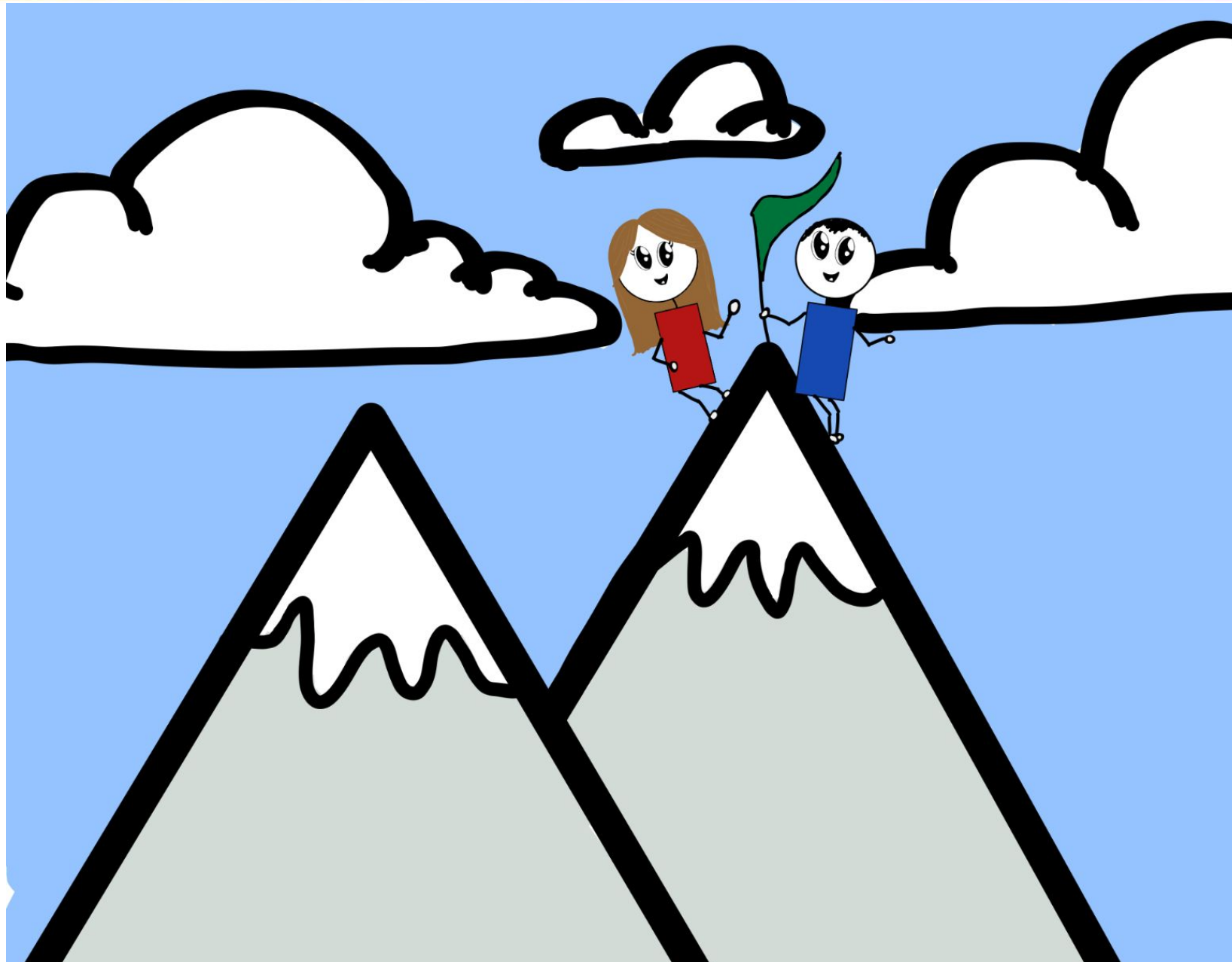


KubeCon



CloudNativeCon

Europe 2025



# End of the Journey (or is it?)



KubeCon



CloudNativeCon

Europe 2025





KubeCon



CloudNativeCon

Europe 2025

# We are hiring: [bloomberg.com/engineering](https://bloomberg.com/engineering)

- Gen AI Platform Team Lead - AI Engineering (London)
- Senior Software Engineer - Data Technologies Compute Platform (London)
- Senior Software Engineer - AI Hardware (NYC)
- Senior ML Ops Engineer - AI Engineering (NYC)
- Senior Software Engineer - Bare Metal as a Service (NYC)
- Senior Software Engineer - Public Cloud IAM and Org Management (NYC)
- Senior Software Engineer - Public Cloud Pipelines (NYC)
- Senior Software Engineer - Public Cloud Visibility (NYC)
- Senior Java Engineer - Search Infrastructure (NYC)
- Technical Product Manager, GenAI Developer Platform - CTO Office (NYC)
- Technical Product Manager, LLM Platforms - CTO Office (NYC)

# Bloomberg

Contact Our Recruiters



**TechAtBloomberg.com**

© 2025 Bloomberg Finance L.P. All rights reserved.



KubeCon



CloudNativeCon

Europe 2025

We are hiring: <https://join.com/companies/liquidreply>

- Business Unit Manager - Sovereign Cloud - Germany
- Developer Relations Engineer - Platform Engineering, Open Source & Sovereign Cloud - Germany
- Platform Engineer - Germany
- Cloud Engineer - Germany

Grab your free copy with signing of  
**“Platform Engineer for Architects - Crafting modern platforms as a product”**

- Today - Booth S240 @Dynatrace during booth crawl
- Tomorrow - Booth S641@Syntasso during lunch break

