From IDP to AiDP: Evolving Your Platform for the Machine Learning Age

Max Körbächer - Liquid Reply

AI: The New Frontier

1. **Past**

   Decade of debate on K8s suitability for stateful workload, DBs and others

2. **Present**

   Platforms are evolving for AI development, integration, and execution

3. **Future**

   AI-native environments must be able to adjust to even more diverse workload

Does It Need to Be Kubernetes?

No.

Bare Metal

Direct GPU access, higher performance

Notebooks

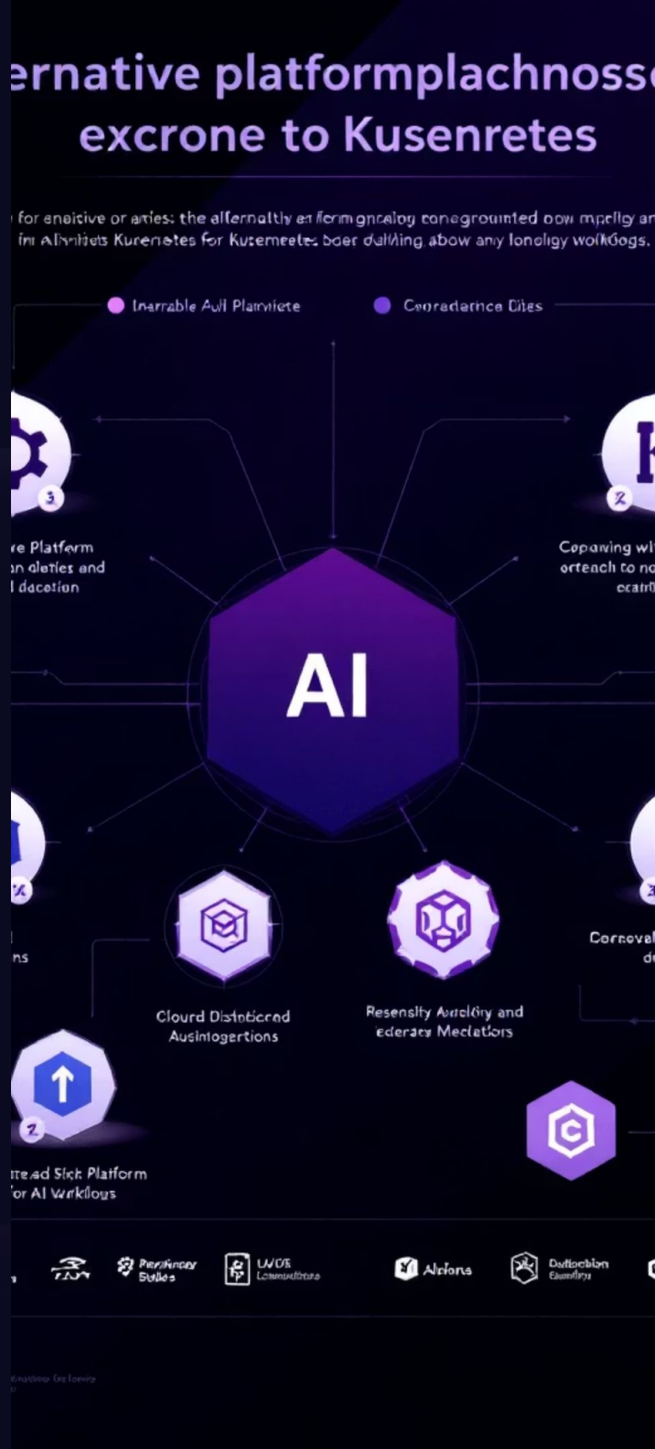
Preferred by researchers

Specialized ML Platforms

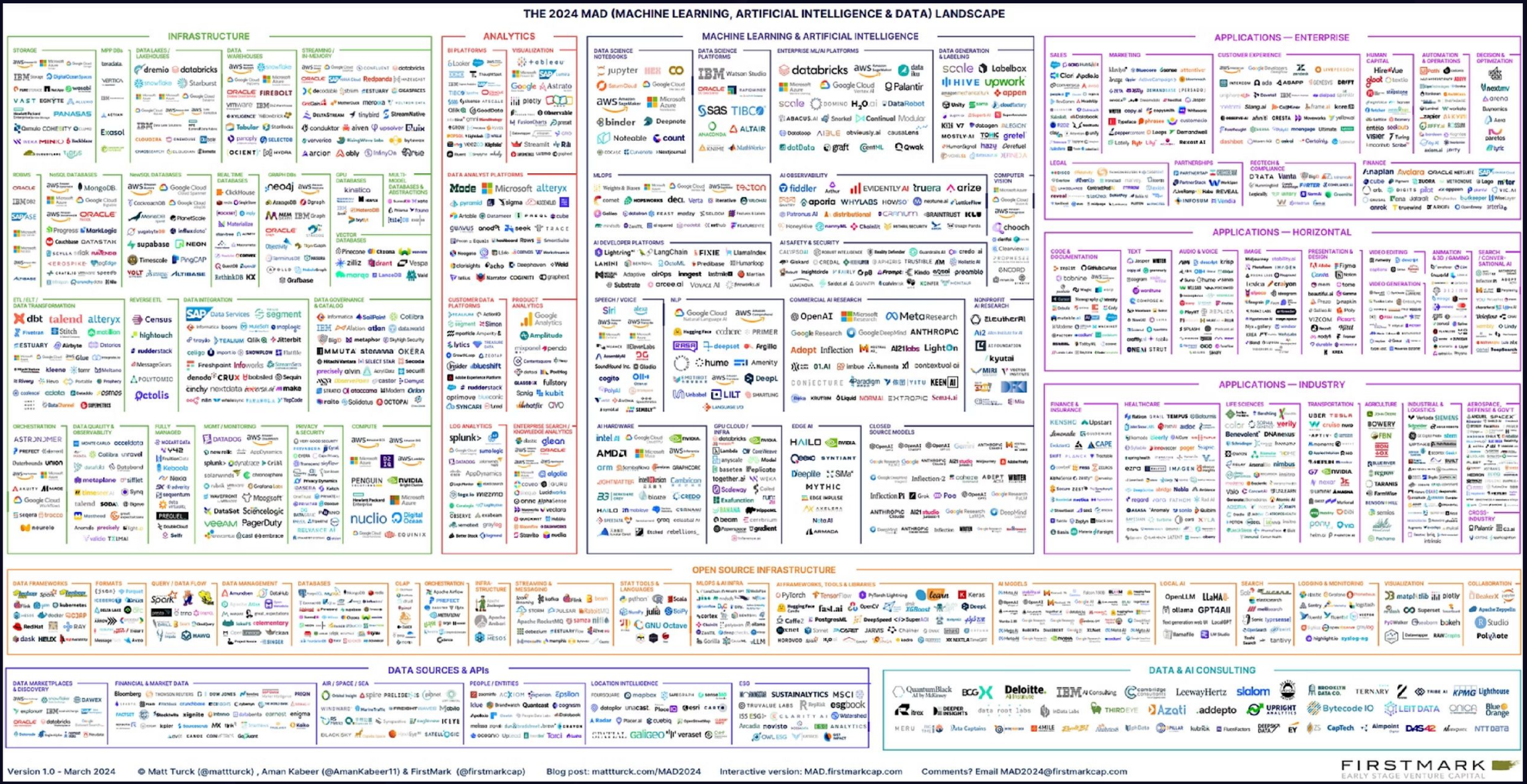
Built for AI workflows

Cloud AI Services

Managed solutions, less overhead



THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE



Bridging the Gap: Dev and Research Teams

Traditional Developers

Code-first approach

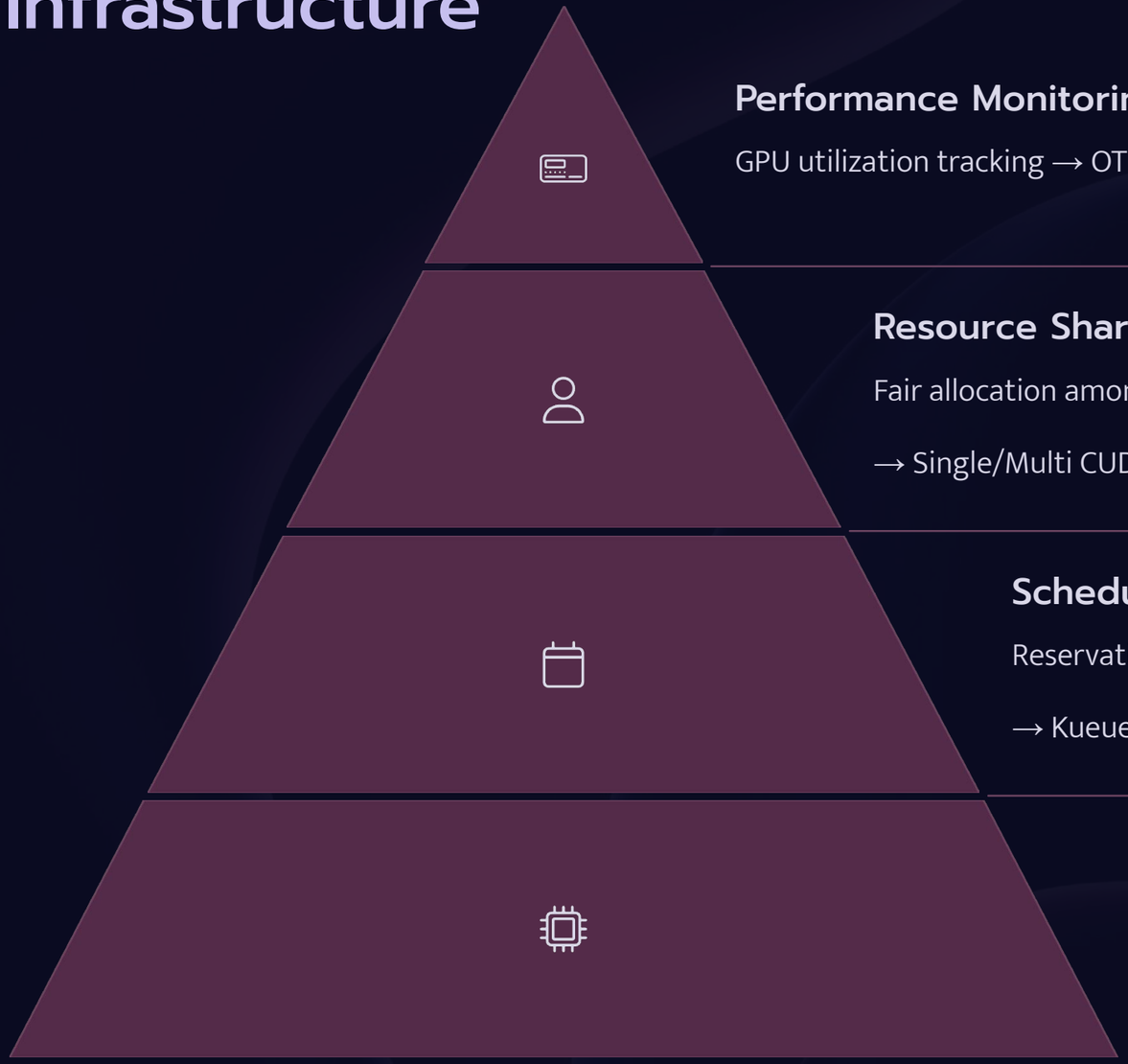
- Small artifacts, predictable CI/CD, auto. testing
- Standard hardware resources
- Scale horizontally

AI/ML Researchers

Experiment-first approach

- Iterative experimentation, complex flows
- Giant models, massive datasets
- Specialized hardware requirements
- Scale vertically, mix

The Foundation: Specialized Infrastructure



Performance Monitoring

GPU utilization tracking → OTEL + Observability Queries

Resource Sharing

Fair allocation among teams

→ Single/Multi CUDA, Time-Slicing, MIG, vGPU, DRA, KAI

Scheduling

Reservation systems for GPUs

→ Kueue, Kubeflow, Yunikorn, Armada, Volcano

Hardware Access

GPUs, TPUs, specialized accelerators

Managing the Lifeblood: Data

Storage

High-performance datastores

Versioning

Track dataset changes

Distribution

Efficient delivery to compute

Processing

Transform and clean data

Following the Hype: GenAI and AI Inferencing

Model Training

Resource-intensive, batch processing

Inference

Optimized for speed, low latency

LLM Integration

API facades, prompt engineering

Deployment Patterns

From embedded to API services

TensorFlow

vLLM

Streamlit

Models as First-Class Citizens - same but different

Development

Create model architectures

Training

Run on specialized hardware

Evaluation

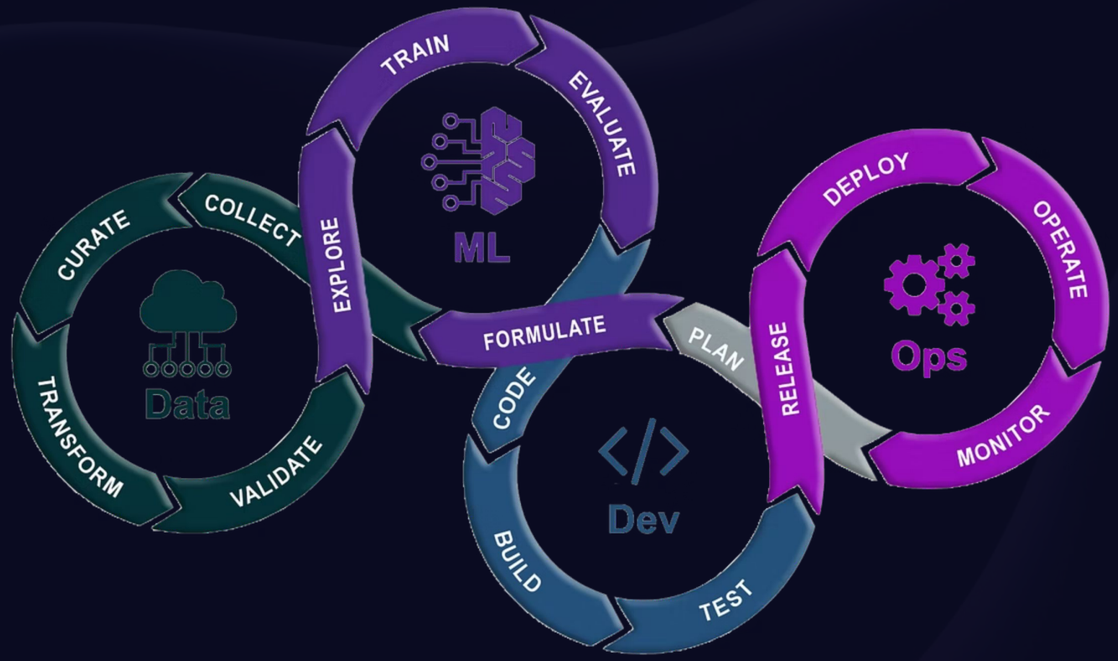
Test against metrics

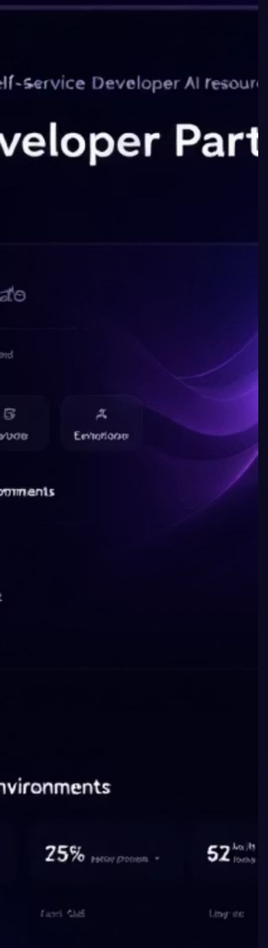
Registry

Store with metadata

Deployment

Serve to production





Democratizing AI

Self-Service Portal

- GPU quota management
- On-demand environments
- Cost visibility
- Simple Inferencing and Dependency Management

Environment Templates

- Pre-configured ML stacks
- Notebook environments
- Training clusters

Tool Integration

- Popular ML frameworks
- Experiment tracking
- Model management

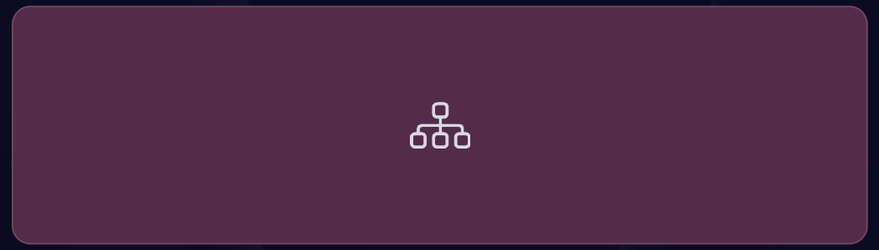
Evolving Your Platform: A Roadmap

Foundation

Basic GPU access, data storage, advanced scheduling

Enablement

ML tooling, training support, simple inferencing



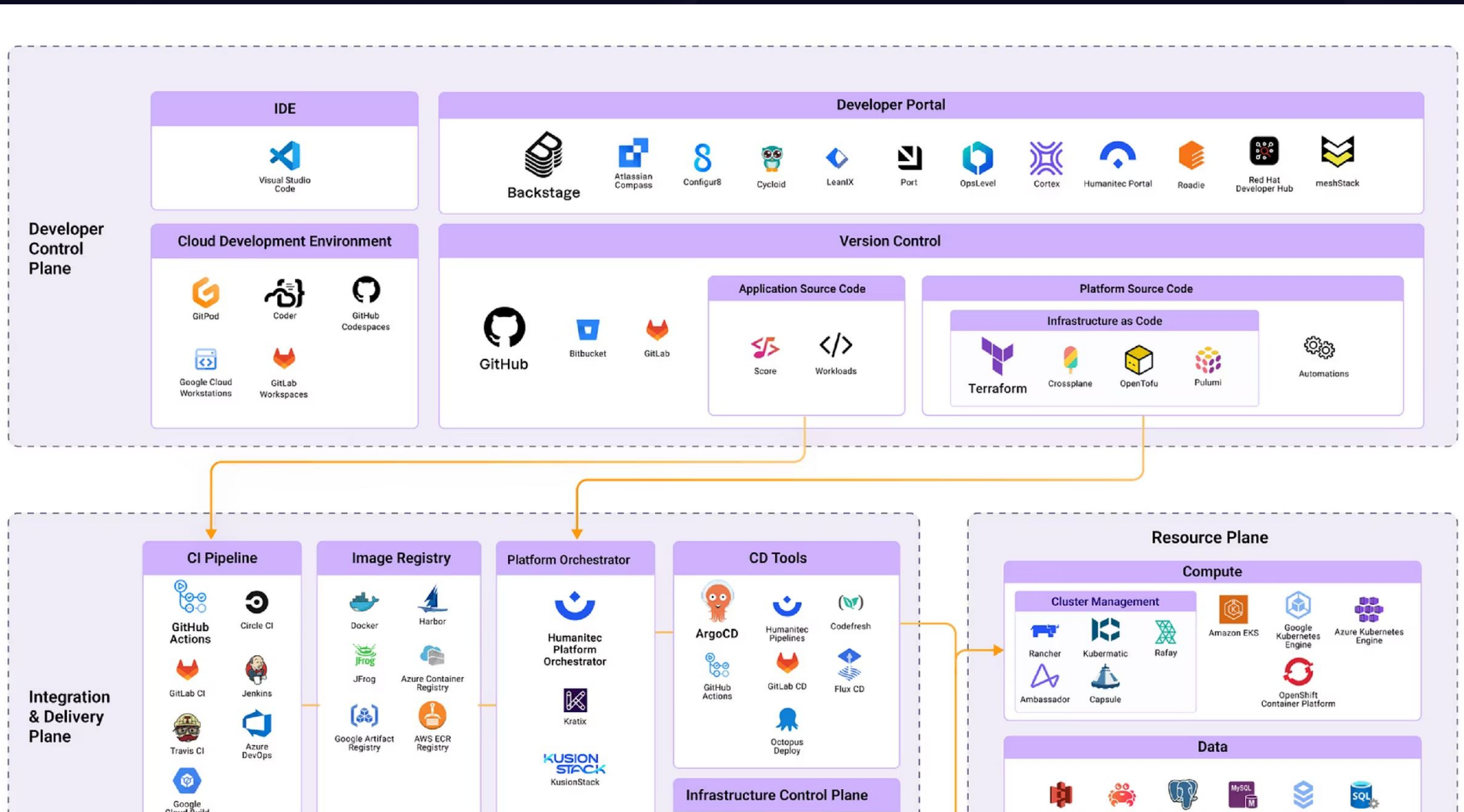
Integration

CI/CD for models, monitoring, registries

Innovation

AI API Gateways, MCP server

A Reference Architecture is not enough



Beyond IDP: The Complex AI Ecosystem

Complex Ecosystem

AI architectures are facing a complex ecosystem, where a reference architecture often follows a single choice of tech.

Integration Landscape

New ML tools and services continuously emerge, demanding flexible integration patterns.

External Dependencies

Model registries, data catalogs, and specialized compute providers create a complex dependency graph.

AI-enabled platforms exist both within and beyond the IDP boundaries. They form intricate ecosystems that must evolve continuously to support emerging AI capabilities.

Measuring the Impact of Your AI-Ready IDP

AI Serving Signals

Token Consumption

Daily processing volume

Guardrail Success

Protection execution rate

PII Leakage

Sensitive data exposure rate

AI Performance Metrics

Inference Latency

Average response time

Availability

System uptime for AI services

Throughput Gain

vs. non-optimized deployment

AiDPs Enable ISO 42001

AI-ready Internal Developer Platforms unlock compliance with emerging AI governance standards.

2023

Standard Release

First international standard for AI management systems.

100%

Governance

Complete oversight of AI assets and processes.

42001

Certification

Demonstrates commitment to responsible AI practices.

Key Takeaways

Connect Development Teams

Create common ground for software and ML engineers

Data, Models & Resources

Focus on ML workflows, not generic capabilities

AiDPs as a Product

Democratize AI integration & development

Iterate and Measure

Implement gradual changes with concrete success metrics

AiDPs can be an Enabler for Regulations

Not just a business empowerment, also a compliance safety net