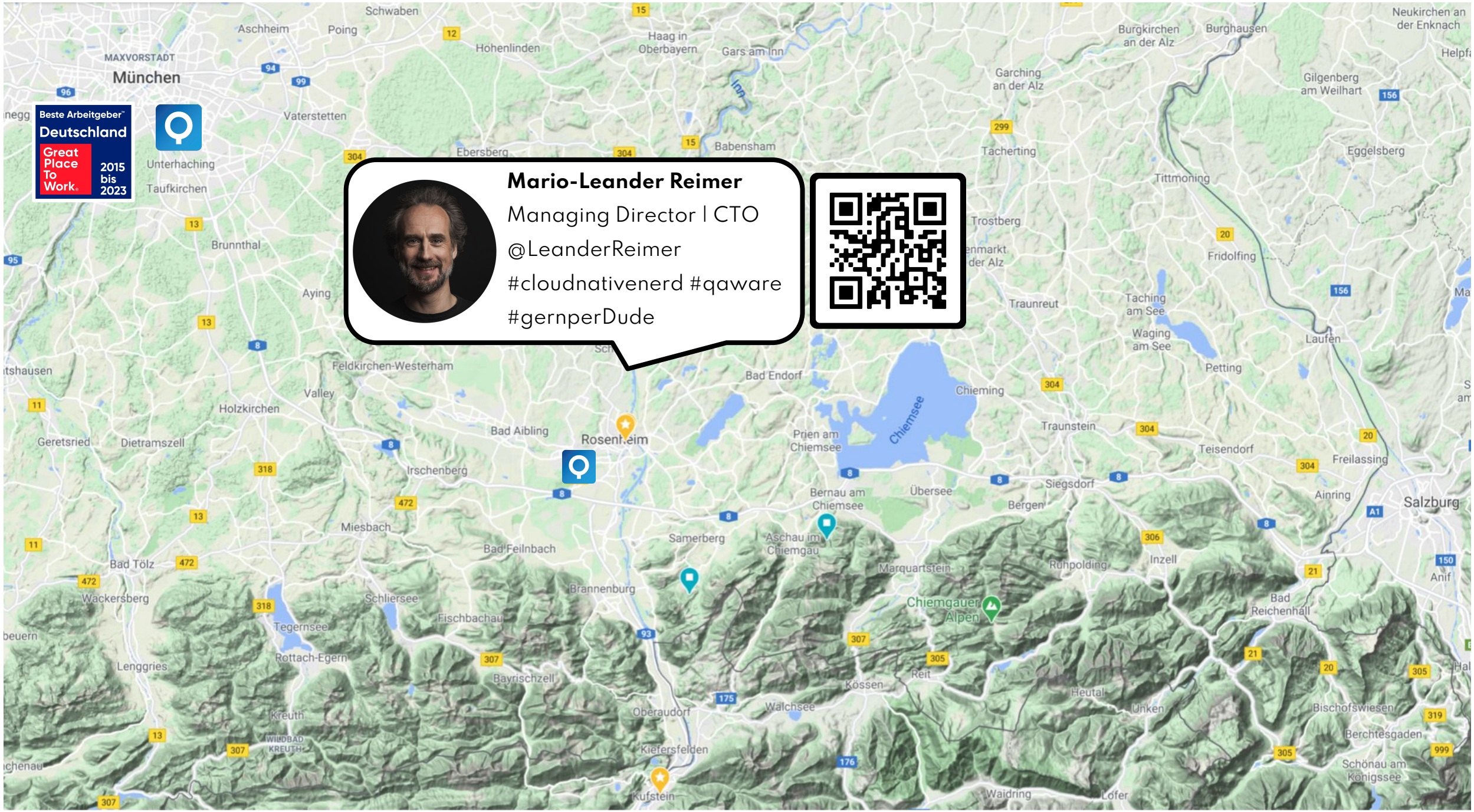
QA|WARE
The Futureware Company

Architecting and Building a K8s-based AI Platform

Mario-Leander Reimer
mario-leander.reimer@qaware.de
@LeanderReimer @qaware
#CloudNativeNerd #gerneperdude


Beste Arbeitgeber Deutschland
Great Place To Work.
2015 bis 2023
Mario-Leander Reimer
Managing Director | CTO
@LeanderReimer
#cloudnativenerd #qaware
#gernperDude
München
Rosenheim
Chiemsee
Salzburg

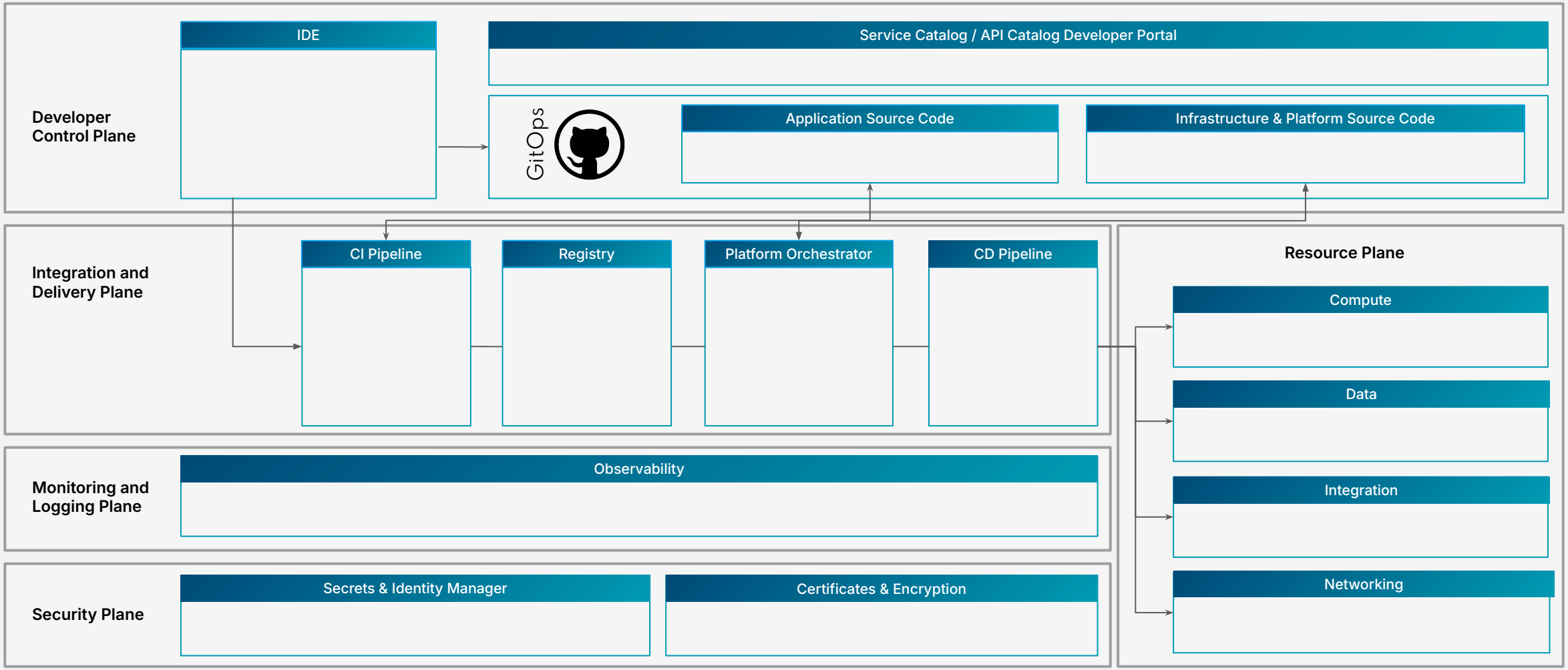
QA|WARE

Platform engineering is the discipline of designing and building toolchains and workflows that enable self-service capabilities for software engineering organizations in the cloud-native era. Platform engineers provide an integrated product most often referred to as an “Internal Developer Platform” covering the operational necessities of the entire lifecycle of an application.

https://platformengineering.org/blog/what-is-platform-engineering

A platform consists of different conceptual components. Depending on the stakeholders and their use cases.

QA|WARE



https://humanitec.com/reference-architectures


QA|WARE
Why do we need an AI platform?

QA|WARE

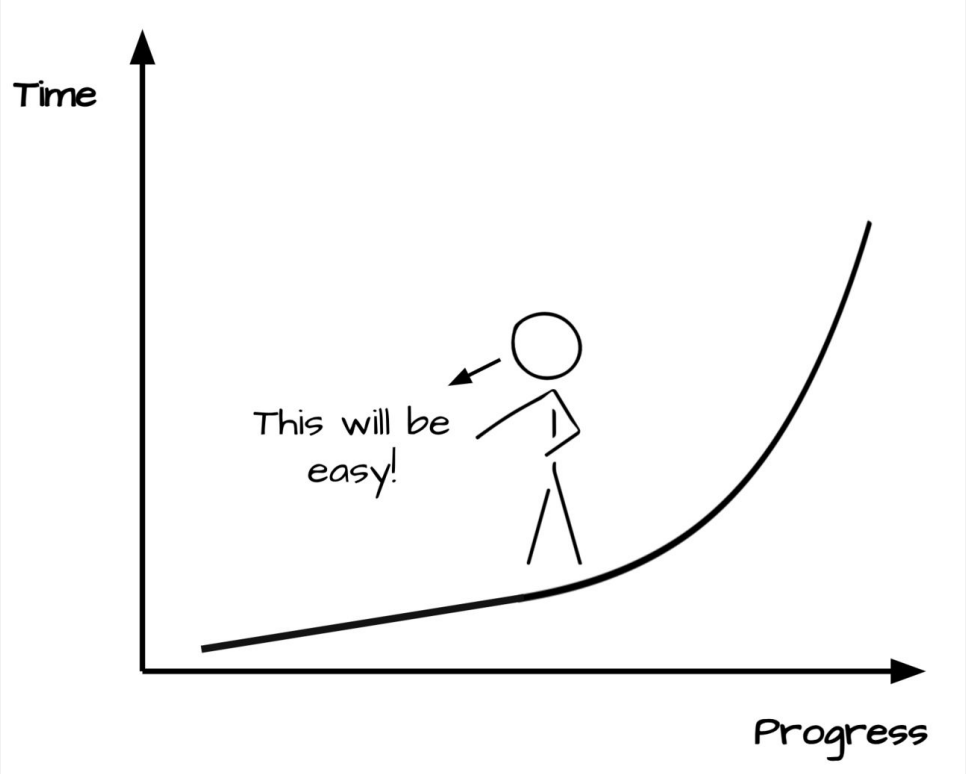
"According to Gartner, 80% of PoCs fail on their way into productive use."

https://www.qaware.de/ki-vom-proof-of-concept-poc-zur-entwicklung/

The 80% Fallacy of AI projects.

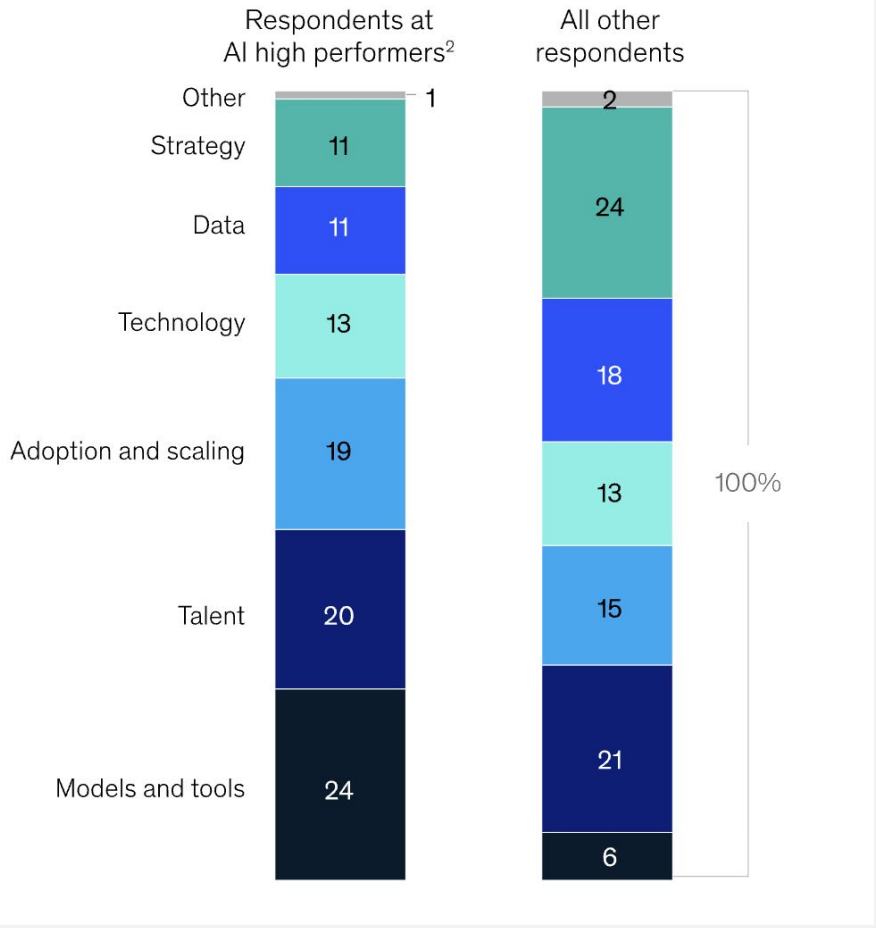
QA|WARE

Tuning ‘routing’ and ‘retrieval’ felt more natural given their classification nature: we built dev sets and fitted them with prompt engineering and in-house models. Now, generation, that was a different story. It followed the 80/20 rule; getting it 80% was fast, but that last 20% took most of our work. When the expectation from your product is that 99%+ of your answers should be great, even using the most advanced models available still requires a lot of work and creativity to gain every 1%.



Juan Pablo Bottaro, LinkedIn Engineering Blog

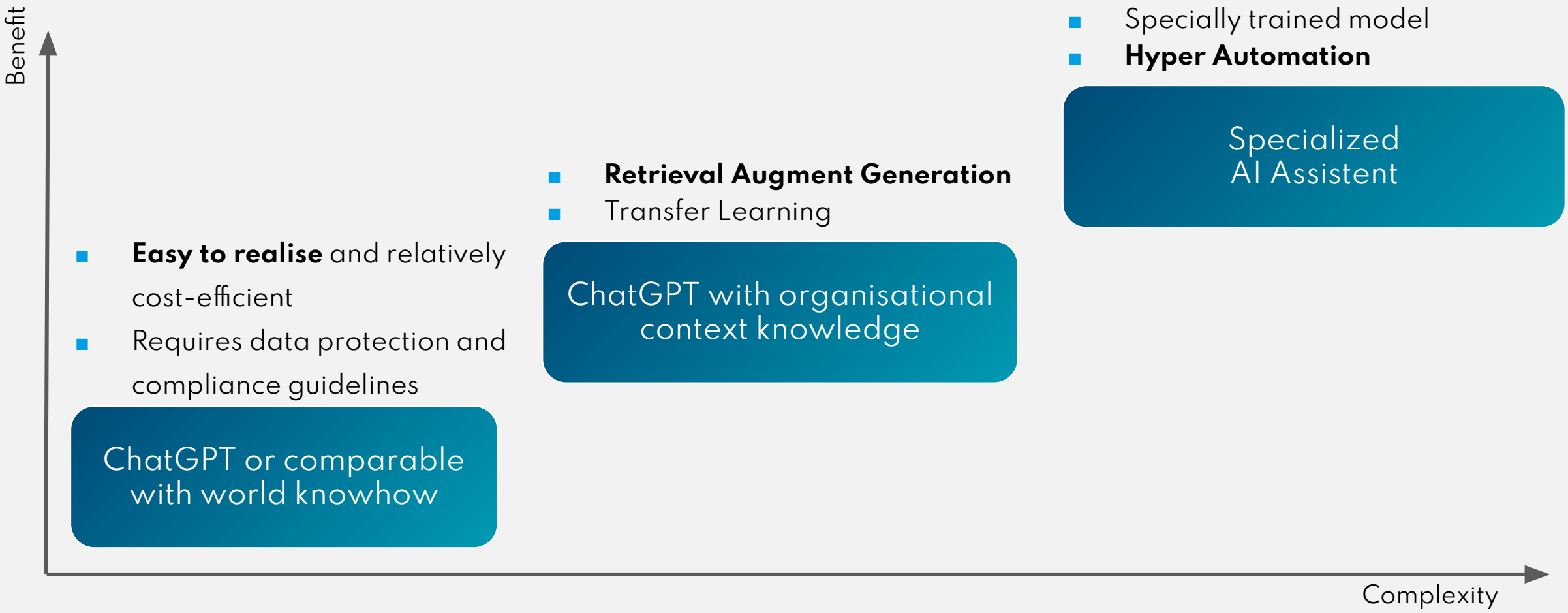
Key challenges: technology, models and tools, scaling.

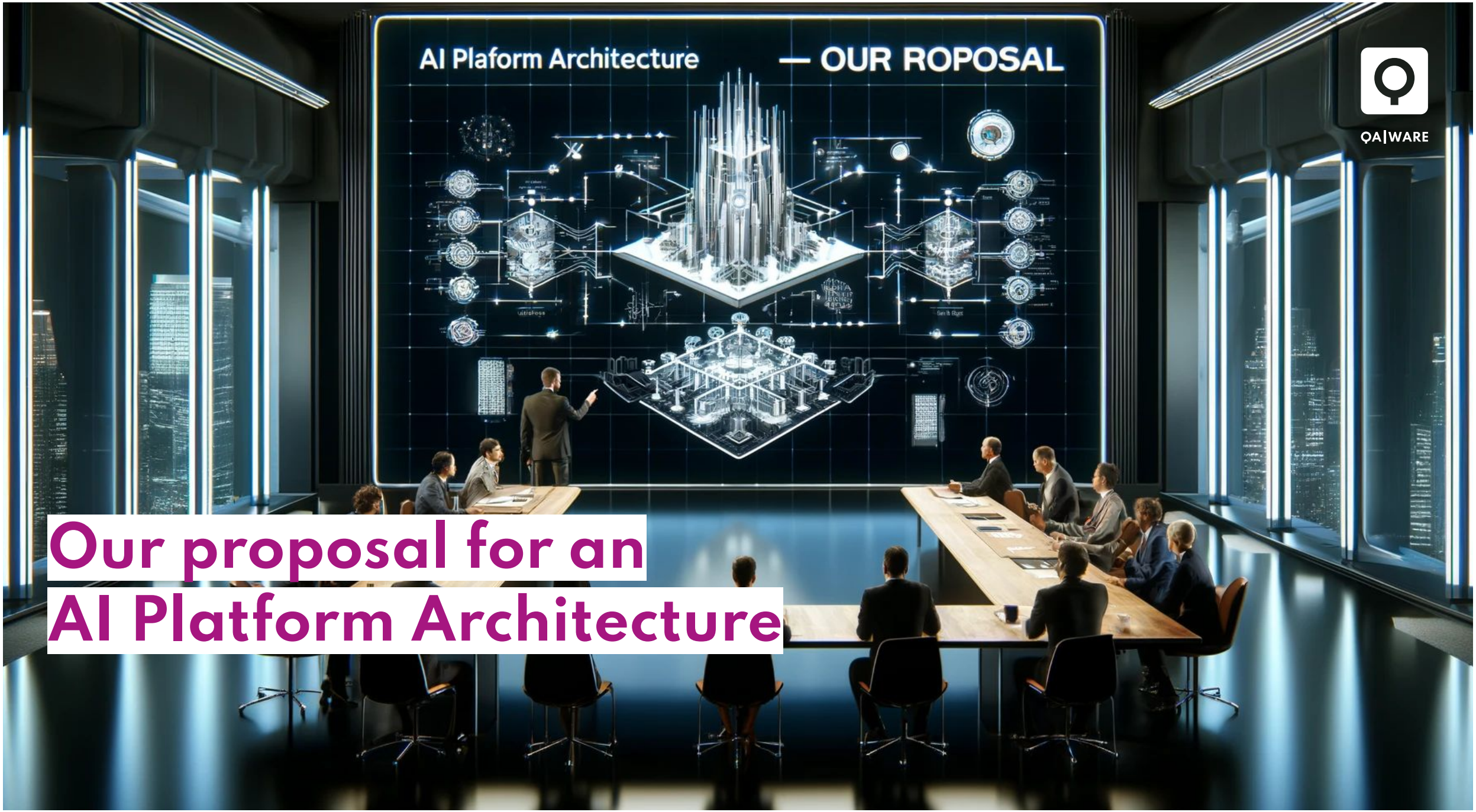


- Different challenges are seen depending on the maturity of the group
- AI newcomers often underestimate the complexity of technologies, models and tools
- Production and scaling challenges often hinder production readiness
- High cognitive load and lack of expertise are also drivers for failing projects

Source: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Chatbots and AI assistants:
The more specific the use case, the more complex it becomes.

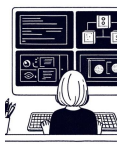




QA|WARE

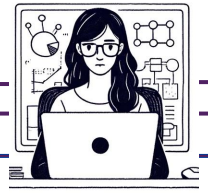
Our proposal for an AI Platform Architecture

Service Plane

- User Serving Plane 
- Access Plane / APIs 
- Orchestration Plane 
- Data Modelling Plane 

Integration & Delivery Plane

- Data Plane
- Model Plane

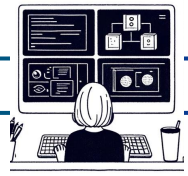


- Quality Plane

- Compliance Plane

Platform Plane

- Observability
- Security
- Delivery
- FinOps
- Operability



Resource Plane

- Compute
- Data
- Integration

QA|WARE

lreimer/k8s-native-ai-platform
lreimer/k3s-ai-platform

The Kubernetes cluster topology requires precise planning. Otherwise the costs will go through the roof!

- There are different GPU machines
- Not all types are available in all regions
- Prices vary drastically, accurate research is recommended
- Additional local SSDs are recommended
- **To be decided:**
  - all nodes with GPU
  - different nodes optimised for normal as well as GPU workloads

```
create-gke-cluster:
    @gcloud container clusters create k8s-native-java-ai \
        --release-channel=regular \
        --cluster-version=1.30 \
        --region=$(GCP_REGION) \
        --addons HttpLoadBalancing,HorizontalPodAutoscaling \
        --workload-pool=$(GCP_PROJECT).svc.id.goog \
        --num-nodes=1 \
        --min-nodes=1 --max-nodes=5 \
        --enable-autoscaling \
        --autoscaling-profile=optimize-utilization \
        --enable-vertical-pod-autoscaling \
        --machine-type=n1-standard-8 \
        --accelerator type=nvidia-tesla-t4,count=1 \
        --local-ssd-count=1 \
        --logging=SYSTEM \
        --monitoring=SYSTEM
```

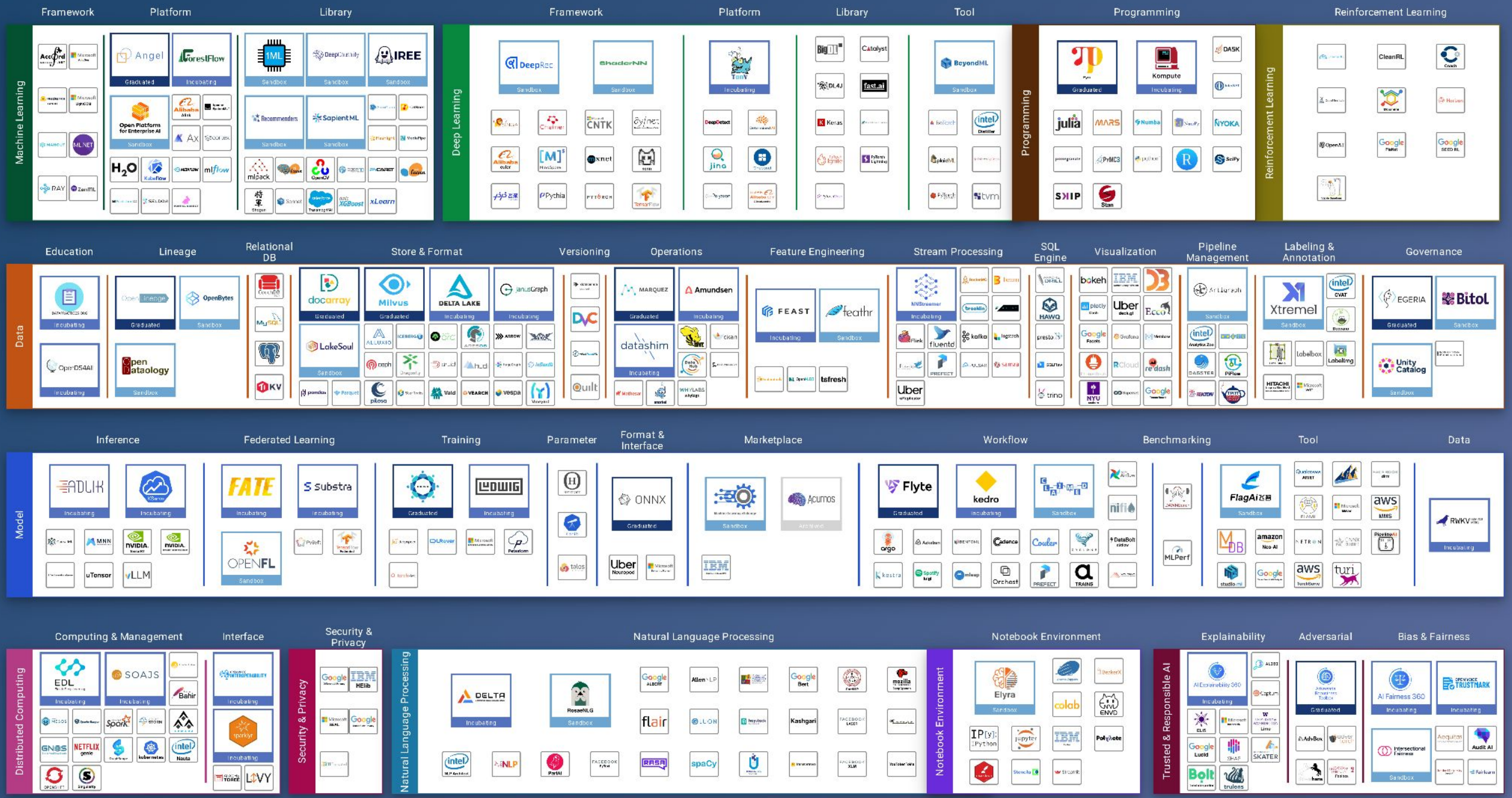
https://cloud.google.com/compute/gpus-pricing?hl=de#other-gpu-models

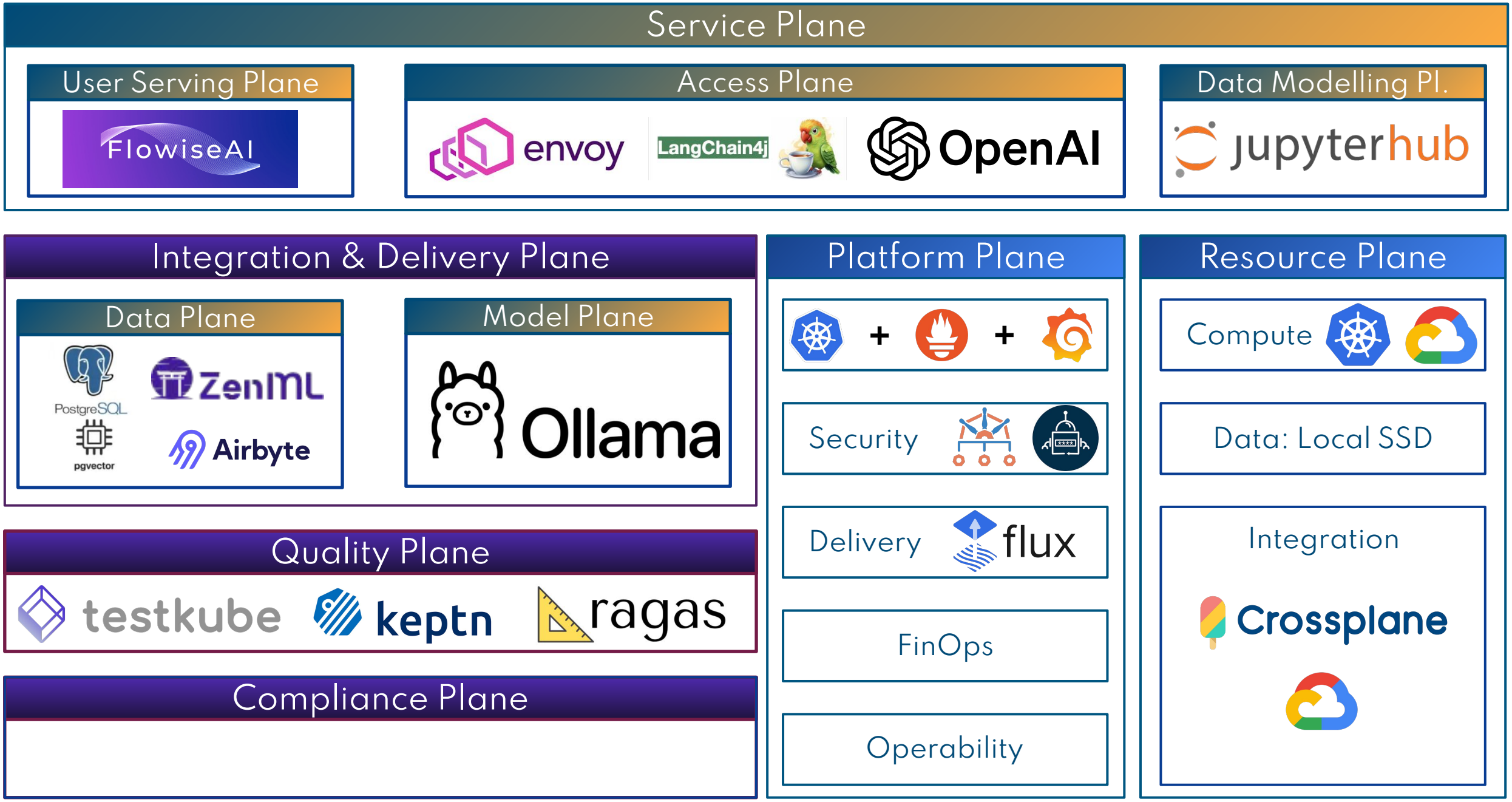
Linux Foundation AI & Data Landscape

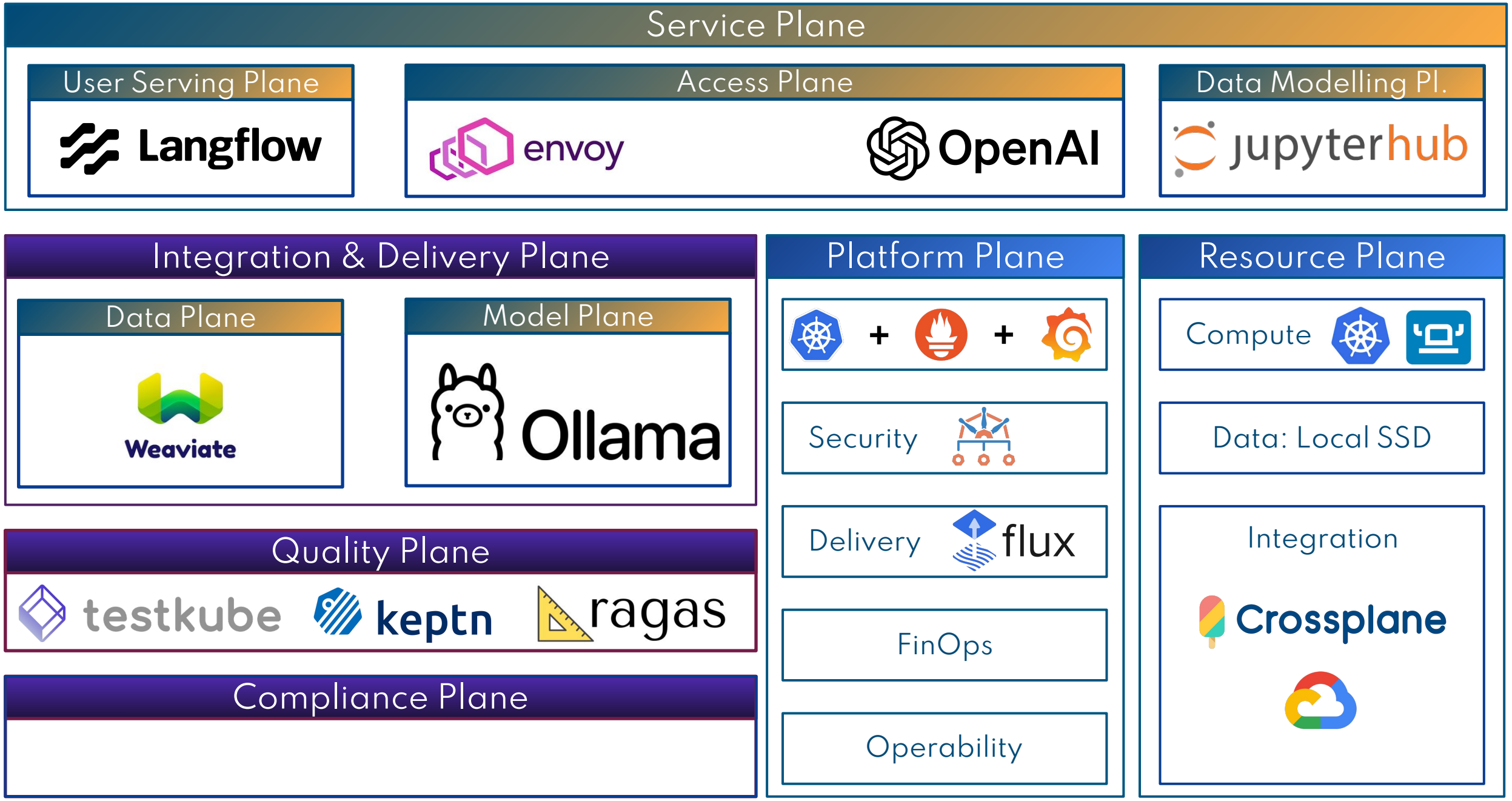
1.0

See the interactive landscape at l.lfaidata.foundation

Greyed logos are not open source




Service Plane
User Serving Plane
FlowiseAI
Access Plane
envoy
LangChain4j
OpenAI
Data Modelling Pl.
jupyterhub
Integration & Delivery Plane
Data Plane
PostgreSQL
pgvector
ZenML
Airbyte
Model Plane
Ollama
Quality Plane
testkube
keptn
ragas
Compliance Plane
Platform Plane
+
+
Security
Delivery
flux
FinOps
Operability
Resource Plane
Compute
Data: Local SSD
Integration
Crossplane


Service Plane
User Serving Plane
Langflow
Access Plane
envoy
OpenAI
Data Modelling Pl.
jupyterhub
Integration & Delivery Plane
Data Plane
Weaviate
Model Plane
Ollama
Quality Plane
testkube
keptn
ragas
Compliance Plane
Platform Plane
+
+
Security
Delivery
flux
FinOps
Operability
Resource Plane
Compute
Data: Local SSD
Integration
Crossplane

QAware GmbH | Aschauer Straße 30 | 81549 München | GF: Dr. Josef Adersberger, Michael Stehnken, Michael Rohleder, Mario-Leander Reimer
Niederlassungen in München, Mainz, Rosenheim, Darmstadt | +49 89 232315-0 | info@qaware.de