



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



Unterhaching
Taufkirchen



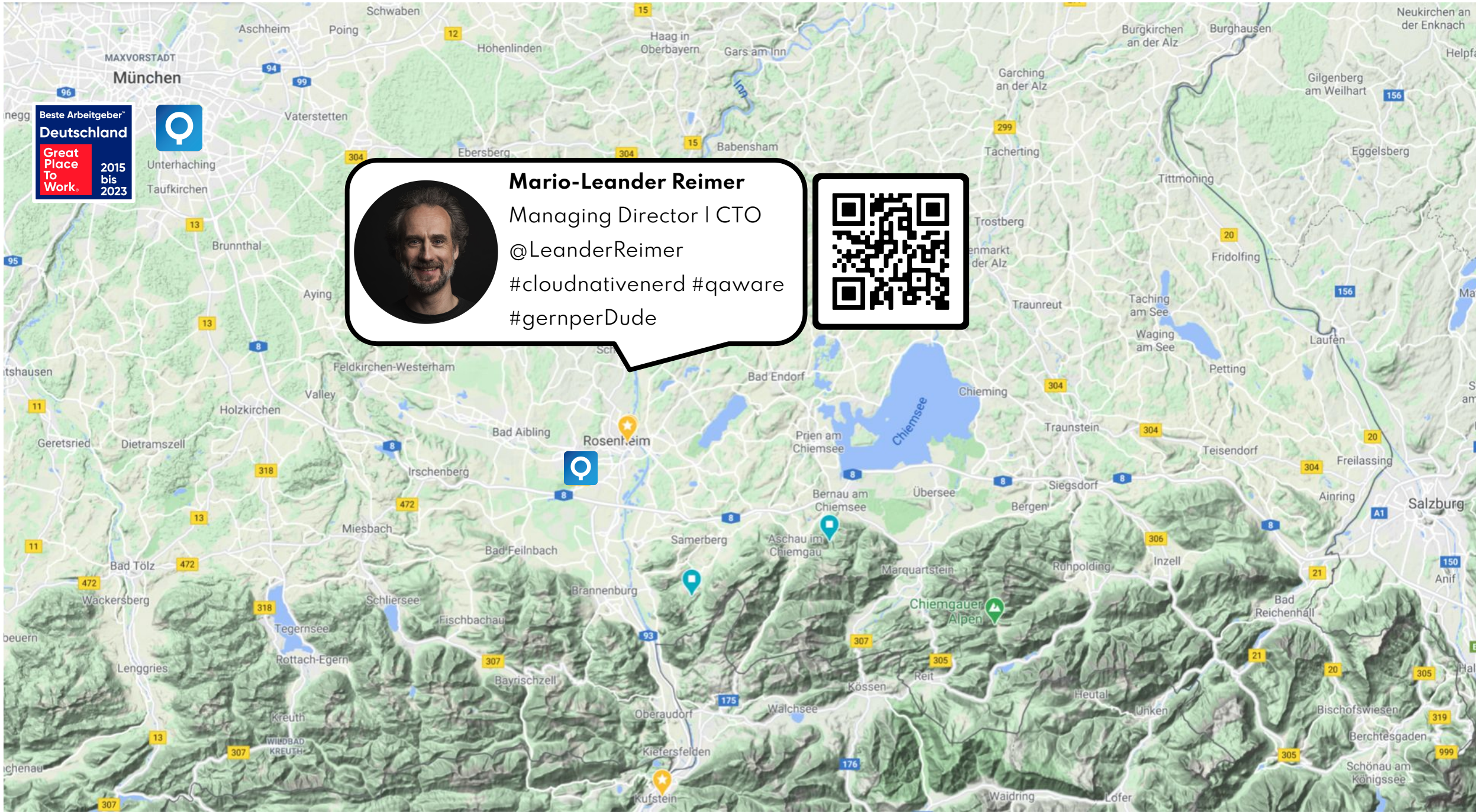
Mario-Leander Reimer

Managing Director | CTO

@LeanderReimer

#cloudbnativenerd #qaware

#gernperDude





QAWARE

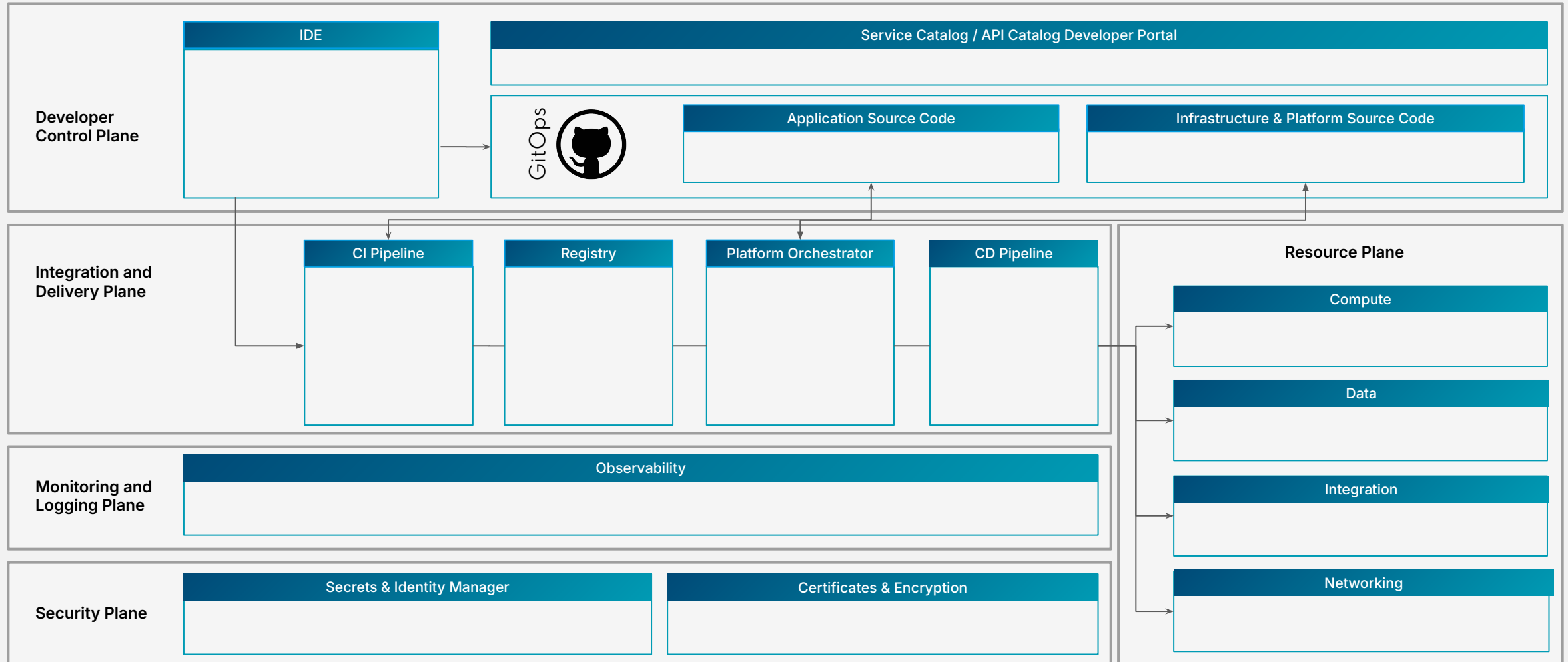
Platform engineering is the discipline of designing and building toolchains and workflows that enable self-service capabilities for software engineering organizations in the cloud-native era. Platform engineers provide an integrated product most often referred to as an “Internal Developer Platform” covering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





QA|WARE



Why do we need an AI platform?



Q|WARE

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

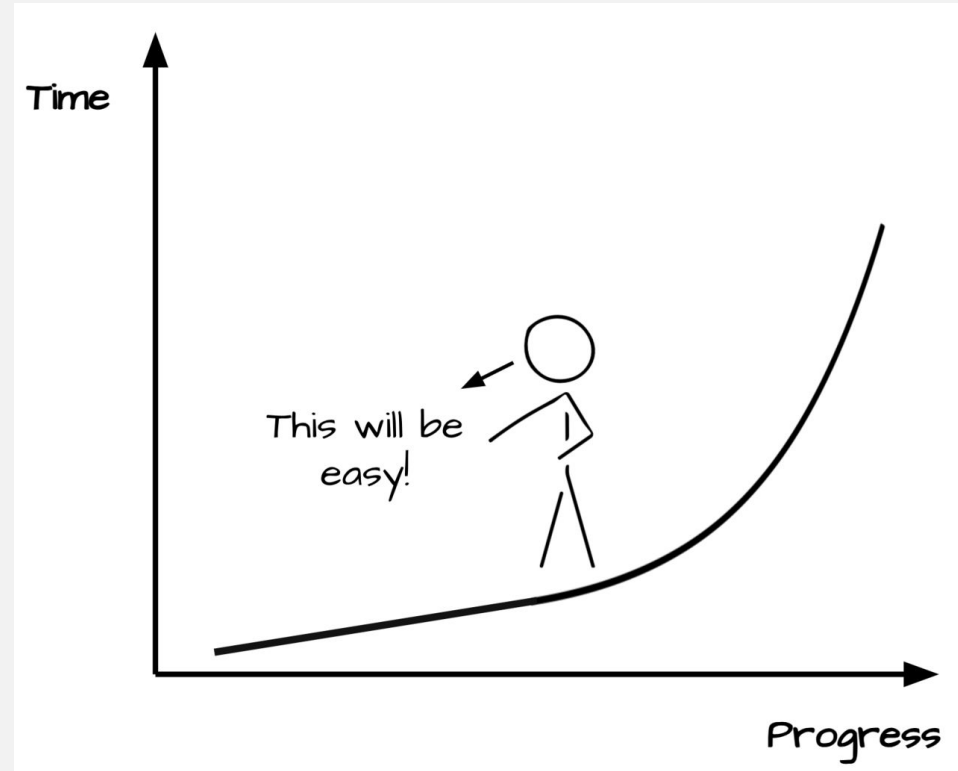
<https://www.qaware.de/ki-vom-proof-of-concept-poc-zur-entwicklung/>

The 80% Fallacy of AI projects.



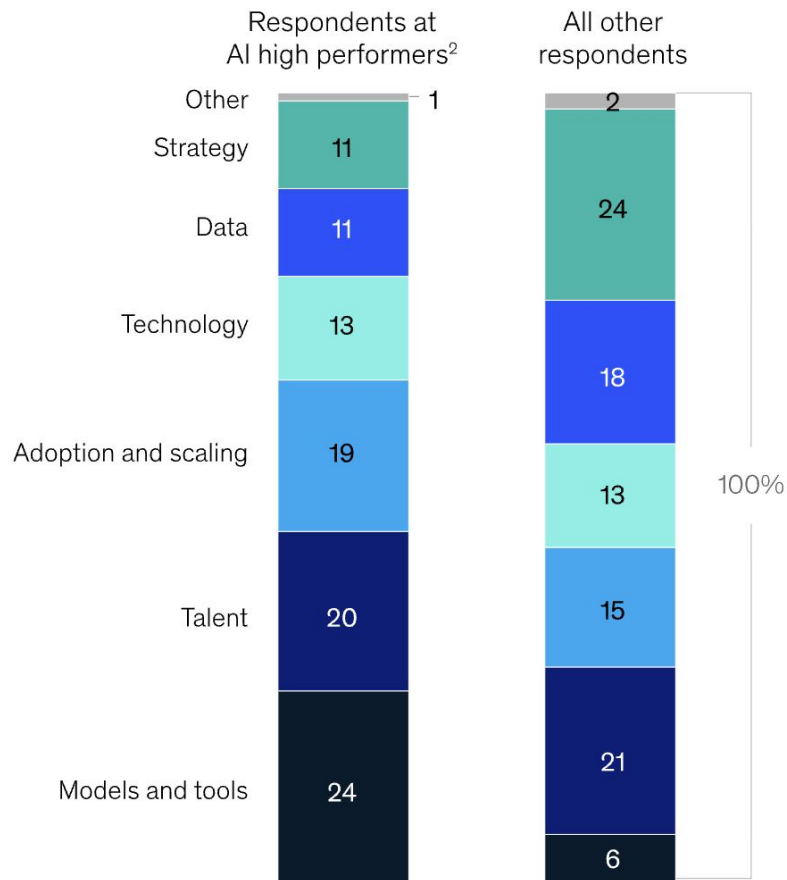
QA|WARE

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%.



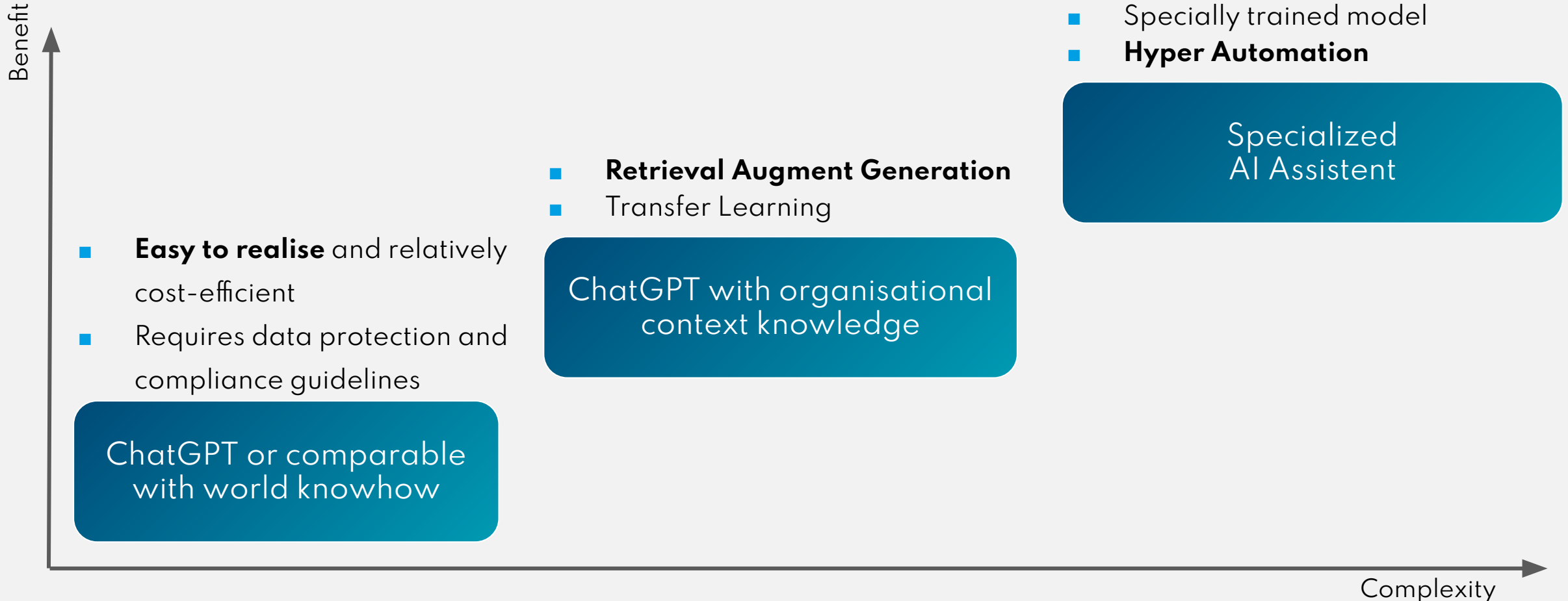
[Juan Pablo Bottaro, LinkedIn Engineering Blog](#)

Key challenges: technology, models and tools, scaling.



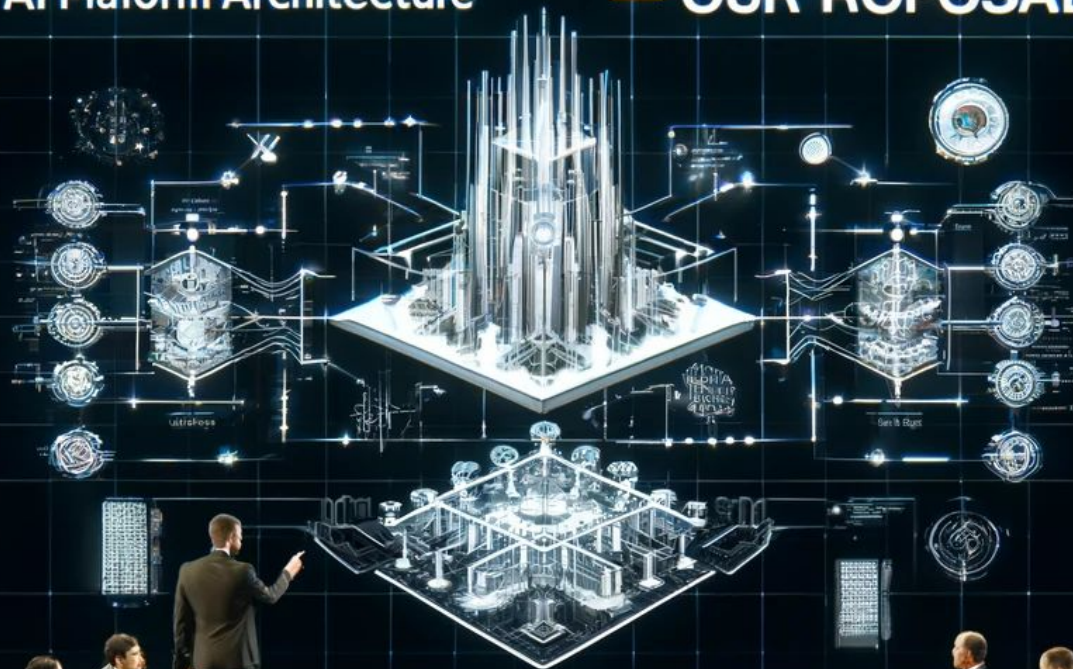
- Different challenges are seen depending on the maturity of the group
- AI newcomers often underestimate the complexity of technologies, models and tools
- Production and scaling challenges often hinder production readiness
- High cognitive load and lack of expertise are also drivers for failing projects

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



AI Platform Architecture

— OUR PROPOSAL



QAWARE

**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

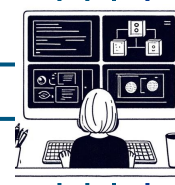
Operability

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!



QA|WARE

- 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>

Machine Learning

Framework	Platform	Library	Framework	Platform	Library	Tool	Programming	Reinforcement Learning
Accord.NET, Microsoft Azure ML, TensorFlow, Microsoft Cognitive Toolkit, PyTorch, ML.NET, RAY, ZenML	Angel, ForestFlow, Open Platform for Enterprise AI, H2O, KubeFlow, mlflow	Intel IML, DeepClarity, IREE, Recommenders, SopleML, mlpack, OpenCV, PyCaret, XGBoost, XLearn	DeepRec, ShaderNN	TOBY	BigDL, Catalyst, DL4J, fast.ai, Keras, TensorFlow, PyTorch, PyTorch Lightning, PyTorch, TensorFlow	BeyondML, Intel Distiller, Intel OneAPI	TP, Kompute, DASK, Julia, MARS, Numba, SciPy, Kiyoka, R, SciPy, SKIP, Stan	CleanRL, OpenAI, Google FALCON, Google SEED RL

Data

Education	Lineage	Relational DB	Store & Format	Versioning	Operations	Feature Engineering	Stream Processing	SQL Engine	Visualization	Pipeline Management	Labeling & Annotation	Governance
OpenDS4AI	OpenLineage, OpenBytes, OpenDataCatalog	MySQL, KV	docarray, Milvus, Delta Lake, JanusGraph, LakeSoul, ALUXIO, ICEBERG, DORIS, ARROW, Hudi, Trino, Valid, YEAH, VESPA, Muxet	DVC, Outlit	Marquez, Amundsen, DASHIM, Whylabs Atlas	FEAST, Featathr	NBStream, Flink, fluento, kafka, logstash, ELK, ELK, ELK, ELK	DRILL, HAWQ, Presto, Databricks, trino	bakeh, Uber, Ecco, Google, RCloud, reDash	Artifactory, DASETE, FISHER	Xtremel, Intel CVAT, Labelbox, Labeling, HITACHI	EGERIA, Bitol, Unity Catalog

Model

Inference	Federated Learning	Training	Parameter	Format & Interface	Marketplace	Workflow	Benchmarking	Tool	Data
ADUIK, MNN, NVIDIA, uTensor, vLLM	FATE, Substra, OPENFL	LEOWIG, ONNX, TensorFlow, PyTorch, ONNX, TensorFlow, PyTorch	ONNX, TensorFlow, PyTorch	ONNX, TensorFlow, PyTorch	Acumos	Flyte, kedro, Argo, Airflow, Prefect, Trainers	MLPerf	FlagAI, Amazon, AWS, Turi	RWKV

Trusted & Responsible AI

Computing & Management	Interface	Security & Privacy	Natural Language Processing	Notebook Environment	Explainability	Adversarial	Bias & Fairness
EDL, SOAJS, NetfliX, Intel, Spark, Storm, Hadoop, Intel, Nucleo	Centralized, TORRENT, TORRENT	Security & Privacy, Google, Microsoft, Intel	DELTA, RoseNLG, Google ALBERT, Allen-LP, flair, LON, Kashgari, Facebook LASEL, spaCy, Pytorch, Facebook XLM, Pytorch	Elyra, Colab, Jupyter, Polynote, IPython, Jupyter, Polynote	AI Explainability 360, CLUE, Google Luceid, Bolt	Adversarial, Adversarial, Adversarial	AI Fairness 360, TRUSTMARK, Intersectional Fairness, Audit AI

Service Plane

User Serving Plane

FlowiseAI

Access Plane



LangChain4j

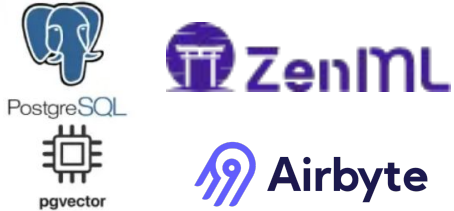


Data Modelling Pl.



Integration & Delivery Plane

Data Plane



Model Plane



Platform Plane



Security



Delivery



FinOps

Operability

Resource Plane

Compute



Data: Local SSD

Integration

Crossplane



Quality Plane



Compliance Plane

Service Plane

User Serving Plane



Access Plane



Data Modelling Pl.



Integration & Delivery Plane

Data Plane



Model Plane



Platform Plane



FinOps

Operability

Resource Plane



Data: Local SSD

Integration



Quality Plane



Compliance Plane



QA|WARE
SOFTWARE ENGINEERING

Thank you!

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