KI-Hosting „Made in Germany“

1. Warum KI-Hosting zur Governance-Frage wird
2. Die Technik dahinter – was produktives KI-Hosting wirklich braucht
3. Hosting-Modelle im Vergleich und der Entscheidungsbaum
4. Referenzarchitektur – SpaceKI
5. Der 12-Monats-Pilotplan – erste Schritte konkret
6. Q&A

Referenten

Maja von Bomhard
Director of Collaboration & AI

SpaceNet AG

Dominik Fabianke
Partner Account Manager | CSP

Dell Technologies

1. Warum KI-Hosting zur Governance-Frage wird

- **KI als kritische Infrastruktur:**
  Was sich 2025/26 grundlegend verändert hat
- **EU AI Act, DSGVO, NIS2:**
  Die drei Regelwerke im Zusammenspiel
- **Die fünf Fragen, die jedes Unternehmen jetzt beantworten muss:**
  - Datenlokation?
  - Schlüsselhoheit?
  - AVV?
  - Drittlandtransfer?
  - Exit-Strategie?

2. Die Technik dahinter – was produktives KI-Hosting wirklich braucht

- GPU-Infrastruktur für KI-Workloads:
  Training, Fine-Tuning, Inferenz
- Dimensionierung der PowerEdge-Systeme
- Multi-Instance GPU (MIG), NVLink, GPU-Interconnect
- S3-kompatibler Objektspeicher, NVMe-Storage-Tiering und High-Speed-Netzwerk als Plattform-Bausteine
- Technische Architektur als Übersetzung des Workload-Profils

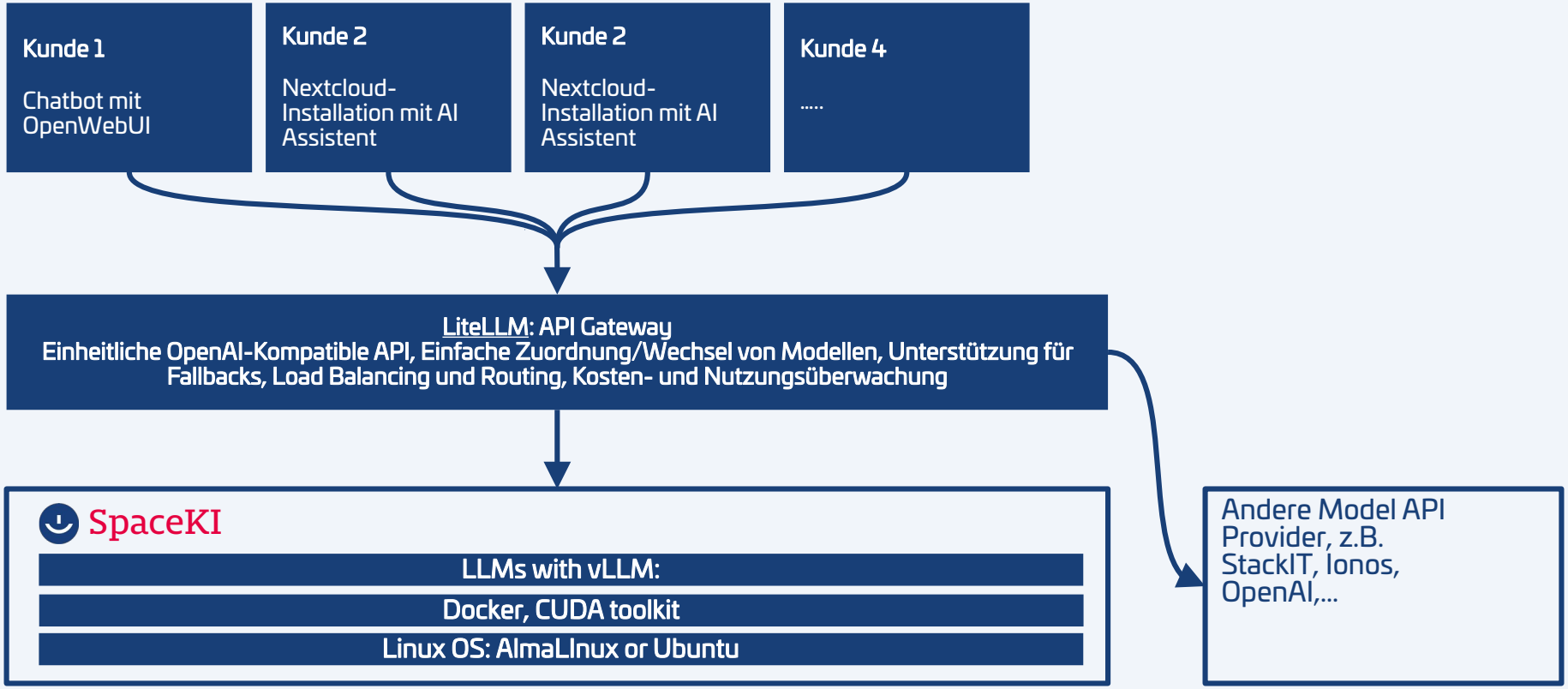
3. Hosting-Modelle im Vergleich & Entscheidungsbaum

- **On-Prem vs. Private Cloud vs. Managed GPU-Hosting (DE) vs. Hyperscaler:**
  ehrliche Gegenüberstellung
- **Hybridstrategien:**
  Wann macht es Sinn verschiedene Workloads auf verschiedene Infrastrukturen zu verteilen?
- **TCO-Realität:**
  Die sechs Kostendimensionen, die viele vergessen

3. Hosting-Modelle im Vergleich

Kriterium	On-Prem	Private Cloud (DE)	Managed GPU (DE)	Hyperscaler
Datenkontrolle	Sehr hoch	Hoch	Mittel-Hoch	Mittel
Time-to-Value	Langsam	Mittel	Schnell	Sehr schnell
GPU-Skalierung	Begrenzt	Gut	Gut	Sehr gut
Drittlandsrisiko	Niedrig	Niedrig	Niedrig-Mittel	Mittel-Hoch
Betriebsaufwand	Hoch	Mittel-Hoch	Mittel	Niedrig-Mittel
Lock-in-Risiko	Niedrig	Mittel	Mittel	Hoch
CAPEX	Hoch	Mittel	Niedrig	Niedrig
Compliance-Klarheit	Hoch	Hoch	Hoch (vertragsabh.)	Komplex

4. Referenzarchitektur SpaceKI – Shared LLM Plattform



4. SpaceKI - Models

Modell	Typ	Größe
gpt-oss-120b	• Large Language Model (LLM) – generativer Text • Liefert enorme Sprachfähigkeiten auf Open-Source-Basis, ideal für große Scale-NLP-Projekte	• 120 Milliarden Parameter; basiert auf einer Transformer-Architektur ähnlich GPT-3/4
ministral-3-14b	• Multimodales Modell (Text + Bild) – „Small-Scale“ Variante des originalen MiniStral • Kompaktes multimodales Modell, das Bild- und Textverarbeitung kombiniert, auch Edge-Geräte	• 14 Milliarden Parameter; Transformer-Encoder-Decoder, kombiniert visuelle und sprachliche Tokens
Qwen-2 1.5B	• Large Language Model (LLM – generativer und instruktions-orientierter Chatbot • Verwendet v.a. für Embedding	• Varianten von 0,5B bis 72B Parameter; Transformer-Architektur von Alibaba Qwen-Team
Whisper	• Automatic Speech Recognition (ASR) – Spracherkennung und Übersetzung • Aktuelle State-of-the-Art-ASR-Modell von OpenAI, das zuverlässig Sprache transkribiert und übersetzt. Dank seiner Open-Source-Verfügbarkeit leicht in Projekte einbindbar	• Modellgrößen von Tiny (≈39M) bis Large-V2 (≈1,5B) Parameter; Convolution-Transformer-Hybrid

5. Der 12-Monats-Pilotplan – erste Schritte konkret

- **Phase 1 (Monat 1):**
  Use-Case-Selektion und Vorbereitung
- **Phase 2 (Monate 2-3):**
  Architektur- und Sicherheitsdesign
- **Phase 3 (Monate 3-5):**
  Proof of Concept und Technischer Pilot
- **Phase 4 (Monate 6-9):**
  MVP in Produktion
- **Phase 5 (Monate 9-12):**
  Skalierung und Portfolio-Ausbau


Wir freuen uns auf
Ihre Fragen!


Vielen Dank für Ihre
Aufmerksamkeit!