SpaceNet

Roadmap zur Implementierung einer KI-Plattform

mit voller Compliance durch deutsches Hosting

PERSÖNLICH.
STARK.
SICHER.
KLUG.
RICHTIG GUT.

Phasenplan

Die folgende Roadmap beschreibt einen 6-12-Monats-Plan für den Aufbau einer produktiven, sicheren KI-Plattform auf deutschem Hosting. Der Plan ist bewusst phasenbasiert, um Risiken zu minimieren, schnell Lernergebnisse zu erzeugen und ein Produktions-Deployment mit belastbaren Compliance-Artefakten zu erreichen.

Phase 1: Use-Case-Selektion und Vorbereitung (Monat 1)

Ziel: Strategische Grundlage und Entscheidungsreife für das Hosting-Modell herstellen.

> Erarbeitung eines Use-Case-Portfolios
(3-5 KI-Kandidaten) mit vorläufiger Risikoeinstufung (Hochrisiko AI Act Annex III ja/nein, DSFA-Indikation nach Art. 35 DSGVO).

> Rollenmapping nach DSGVO
(Verantwortlicher/Auftragsverarbeiter) und AI Act (Provider/Deployer) für jeden Use Case.

> Dateninventur:
Dokumentation, welche Datenquellen in Prompts, RAG-Systeme oder Training einfließen sollen. Klassifizierung nach Sensitivität und regulatorischer Kategorie.

> Provider-Shortlist
(2-3 DE-Anbieter) und Erarbeitung eines standardisierten Due-Diligence-Fragenkatalogs (Rechenzentrumsstandorte, Subunternehmerliste, Zertifizierungen, Schlüsselmanagement, Remote-Zugriff-Politik).

> Rollen und Verantwortlichkeiten:
RACI-Matrix für Projekt und späteren Betrieb.

Phase 2: Architektur- und Sicherheitsdesign (Monate 2-3)

Ziel: Technische Zielarchitektur und Compliance-Rahmen spezifizieren.

> Zielarchitektur-Dokument:
Kubernetes-Cluster-Topologie, GPU-Pool-Design (MIG-Konfiguration), S3-Objektspeicher-Anbindung, Serving-Stack (KServe/Triton oder vLLM), Monitoring-Architektur.

> Security Baseline:
IAM-Design, Netzwerksegmentierung, KMS/HSM-Anbindung, Logging-Konzept (was wird protokolliert, mit welcher Retention, unter welchem Zugriffsschutz).

> Vertrags- und Nachweispaket:
AVV-Entwurf, TOM-Anlage, Subunternehmerliste, Auditrechte-Klausel, Incident-Response-Pflichten.

> SLO-Definition:
Konkrete messbare Service Level Objectives für Latenz, Throughput, Verfügbarkeit und Wiederherstellung.

> Datenschutz-Folgenabschätzung (DSFA)
für Hochrisiko-Anwendungsfälle durchführen oder vorbereiten.

Phase 3: Proof of Concept und Technischer Pilot (Monate 3-5)

Ziel: Technische Machbarkeit und Security-Posture validieren, erste Leistungsmessungen durchführen.

> Minimal-Cluster aufbauen (2-4 GPUs) mit vollständigem Security-Rahmen: RBAC, NetworkPolicies, KMS-Anbindung, Audit-Logging aktiv.

> Model-Serving-Stack integrieren (KServe/Triton), End-to-End-Test vom API-Call bis zur Model-Antwort.

> S3-Kompatibilitätstest: Alle benötigten API-Operationen gegen den DE-Hosting-Objektspeicher testen und dokumentieren.

> RAG-Pipeline (wenn relevant): ETL-Prozess, Index-Build, Retrieval-Test mit Access-Control-Verifikation.

> Security-Tests: Tenant-Isolationstest (simulierter Cross-Tenant-Zugriff), IAM-Fehlkonfigurationstest, Audit-Log-Verifikation, Secret-Scan.

> Benchmark-Ausführung: Latenz, Throughput, GPU-Utilization unter realistischen Last-Profilen dokumentieren.

Phase 4: MVP in Produktion (Monate 6-9)

Ziel: Ersten produktiven Use Case mit vollständigem Governance-Rahmen in Betrieb nehmen.

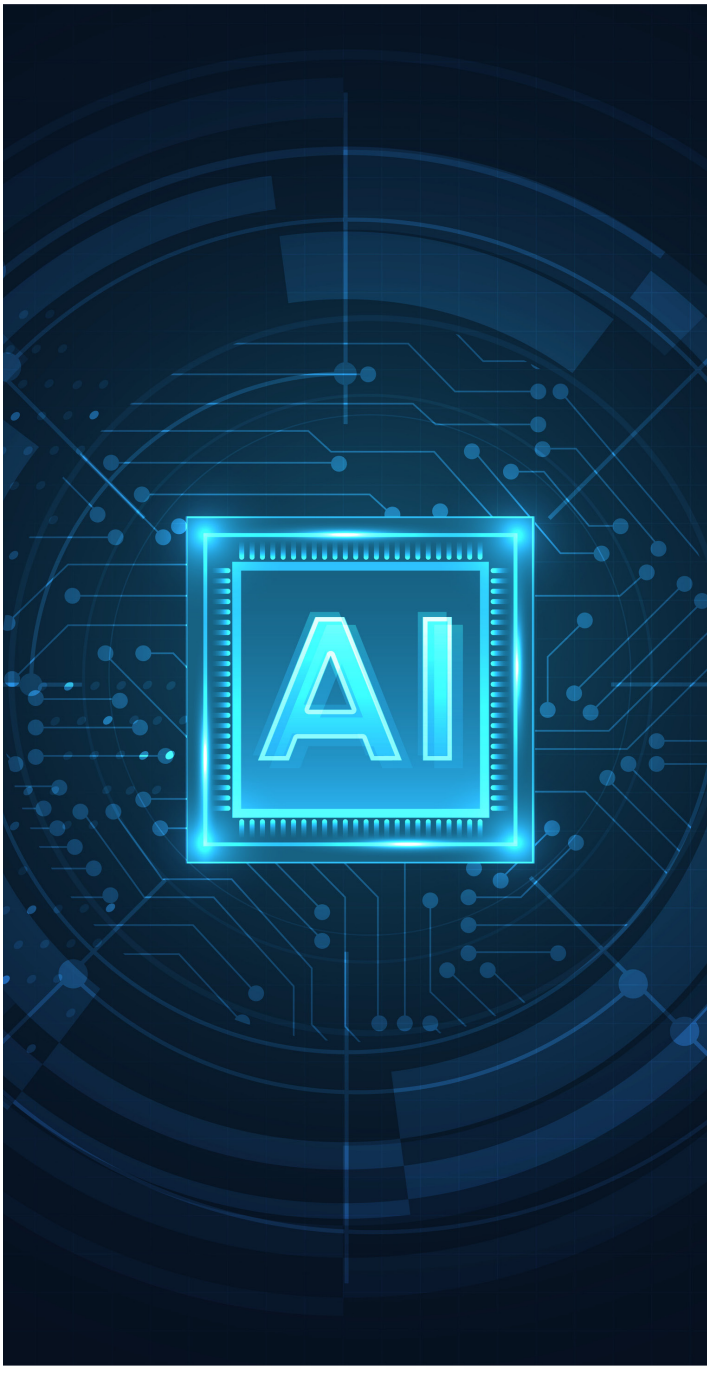
> Ein prioritärer Use Case geht in Produktion (z.B. interner Assistenz-Chat mit RAG auf Wissensdatenbank).

> Governance-Implementierung abgeschlossen: Logging-Retention, Privacy Controls, Rollback-Prozesse, Incident-Response-Runbook.

> SLAs vertraglich vereinbart und monitoring-seitig messbar gemacht.

> AI Literacy Schulungen für betroffene Nutzer und Betriebsteam (Pflicht unter EU AI Act ab 2025).

> Abschluss-Dokumentation audit-ready: Datenflussdiagramm, DSFA, AVV, TOM-Nachweis, Benchmark-Report.



Phase 5: Skalierung und Portfolio-Ausbau (Monate 9-12)

Ziel: Weitere Use Cases onboarden, Plattform für Multi-Tenant-Betrieb optimieren.

> Kapazitätsplanung finalisieren:
Reserved GPUs für kritische Zeitfenster, Multi-Tenant-Strategie mit MIG-Profilen oder dedizierten Node Pools.

> MLOps-Reife steigern:
Supply Chain Signierung für Container Images, Model Registry vollständig integriert, Quality Gates für Modell-Deployments.

> FinOps einführen:
Kosten-Tracking pro Request/Token, Chargeback/Showback für interne Teams, Anomalie-Erkennung bei Kostentreibern.

> Security-Programm verstetigen:
Regelmäßige Penetrationstests, SBOM-Update-Prozess, Vulnerability-Scan-Automatisierung.

Aufgabenbereich	CIO/IT-Leiter	CISO	DPO/Datenschutz	GRC	SRE/Plattform	Data/AI-Team
Hosting-Entscheidung	Verantwortlich	Konsultiert	Konsultiert	Konsultiert	Informiert	Informiert
Provider-Auswahl & AVV	Verantwortlich	Konsultiert	Verantwortlich	Konsultiert	Informiert	Informiert
Security Baseline & Threat Model	Konsultiert	Verantwortlich	Konsultiert	Konsultiert	Mitarbeit	Informiert
DSFA-Ausführung	Konsultiert	Konsultiert	Verantwortlich	Konsultiert	Informiert	Informiert
Plattform-Betrieb & SRE	Konsultiert	Konsultiert	Informiert	Informiert	Verantwortlich	Konsultiert
Modellwahl & KI-Wahl	Informiert	Konsultiert	Konsultiert	Informiert	Konsultiert	Verantwortlich
Compliance-Monitoring	Informiert	Konsultiert	Verantwortlich	Verantwortlich	Konsultiert	Informiert