ÆNIX

Sovereign AI Decision Guide

Sieben Entscheidungen für den Aufbau einer KI-Infrastruktur, die Souveränität, Auditierbarkeit und regulatorische Vorgaben für regulierte Workloads erfüllt.

Für CTOs, Heads of AI/ML & Verantwortliche für Datensouveränität · aenix.io

Dieser Leitfaden führt in fester Reihenfolge durch sieben Entscheidungen für die Architektur einer KI-Infrastruktur für regulierte Workloads. Er kombiniert einen Entscheidungsbaum mit vertiefenden Abschnitten je Entscheidung, vier Referenzarchitekturen und GPU-Sizing-Tabellen für gängige Workloads. Die Zahlen sind grobe Betriebsschätzungen zum Stand des Versionsdatums; validieren Sie sie vor der Kapazitätsfestlegung anhand realer Werte.

Master-Entscheidungsbaum

Durchlaufen Sie die sieben Entscheidungen der Reihe nach. Jede Entscheidung grenzt die Architektur ein; frühere Entscheidungen schränken spätere ein. Die meisten Unternehmensprojekte stoßen auf mehrere Auslöser gleichzeitig — wählen Sie den bindendsten und lagern Sie die übrigen darauf auf.

1. **Auslöser** — Was ist Ihr Auslöser für souveräne KI? Regulierte Daten / Inferenz-Ökonomie / Auditierbarkeitsvorgabe / Air-Gap-Anforderung.
2. **Regulierung** — Welche Regulierung(en) gelten? DORA (Finanzsektor) / NIS2 (wesentliche Einrichtungen) / Sektoral / Nationales Souveränitätsmandat.
3. **Modellklasse** — Open-Weight (Llama, Mistral, Qwen, DeepSeek, Phi, Gemma) oder proprietär (Closed-Weight, herstellerlizenziert)? Souveränitäts- und Auditierbarkeitsanforderungen führen in der Regel zu Open-Weight.
4. **Hardware** — Welche GPU-Klasse? H100/H200 / A100 / L40S / Blackwell (B100/B200) / CPU-only oder kleinere Beschleuniger.
5. **Mandantenfähigkeit** — Single-Tenant / Multi-Tenant über Tenant-CRD (gemeinsamer Cluster, logisch isoliert) / Cluster-pro-Mandant (strikte Isolation).
6. **Souveränität** — Standard (In-Region-Basislinie) / Strikt (kundenkontrollierte Schlüssel, Lieferantentransparenz, Audit-isoliert) / Air-Gapped (offline, Updates nur out-of-band).
7. **Betrieb** — Kundenbetrieben / Anbieterbetrieben (von Aenix verwaltet) / Hybrid (Control Plane beim Anbieter + Data Plane beim Kunden oder umgekehrt).

Die sieben Entscheidungen

Entscheidung 1 — Auslöser

Was ist Ihr primärer Auslöser für souveräne KI? Mehrere können zutreffen; der primäre Auslöser bestimmt die Architektur.

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Regulierte Daten	Dürfen die Rechtsordnung oder die Anbieterumgebung nicht verlassen. Erzwingt Verschlüsselung, Schlüsselverwahrung, Auditierbarkeit.	Regulierte Branche (Finanzwesen, öffentlicher Sektor, Gesundheitswesen). Workload verarbeitet personenbezogene oder geschäftskritische Daten.	Verschiebt Entscheidung 6 (Souveränität) Richtung „Strikt" oder „Air-Gapped".

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Inferenz-Ökonomie	Inferenz im großen Maßstab wird auf eigener Infrastruktur günstiger als per Token über eine API.	Hochvolumige Inferenz (Millionen Tokens/Tag oder dauerhafte Produktionslast). RAG-lastige Anwendungen.	Entscheidung 4 (Hardware) auf Inferenz optimiert (L40S, H100) statt auf Training.
Auditierbarkeitsvorgabe	Vollständige Reproduzierbarkeit nötig — Modell + Daten + Prompt + Antwort unveränderlich protokolliert.	Regulierte Entscheidungsfindung (Kreditvergabe, Versicherungsunderwriting, Verwaltungsabläufe).	Entscheidung 7 (Betrieb) Richtung kundenbetrieben oder striktes Hybrid.
Air-Gap-Anforderung	Workload muss offline laufen — kein Call-home des Anbieters, keine SaaS-Abhängigkeiten.	Verteidigung, souveräne Cloud, isolierte Industrieumgebungen.	Entscheidung 6 auf „Air-Gapped" festgelegt. Entscheidung 3 auf Open-Weight festgelegt.

> Die meisten Unternehmens-KI-Projekte entdecken zwei oder drei Auslöser gleichzeitig. Wählen Sie den bindendsten — jenen, der für sich allein bereits ein souveränes Deployment erzwingen würde. Dieser bestimmt die Architektur; die übrigen werden darauf aufgesetzt.

Entscheidung 2 — Regulierung

Welche Regulierung oder welcher regulatorische Rahmen gilt für Ihren KI-Workload?

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
DORA (Verordnung (EU) 2022/2554)	Betriebliche Resilienz, IKT-Drittparteirisiko, Lieferanten-Exit. Architekturseitig anspruchsvolle Erwartungen.	Regulierte Finanzeinheit oder IKT-Drittdienstleister, der eine solche bedient.	Verschiebt Entscheidung 7 Richtung kundenbetrieben oder vertraglich strikt anbieterbetrieben. Zwingt Entscheidung 5 zu nachweisbarer Isolation.
NIS2 (Richtlinie (EU) 2022/2555)	Risikomanagement, Meldepflichten bei Vorfällen, Lieferkette. Breiter Anwendungsbereich (Energie, Verkehr, Telekommunikation,	Wesentliche oder wichtige Einrichtung gemäß nationaler Umsetzung.	Beeinflusst Entscheidung 6 (Souveränität) und Entscheidung 7 (Betrieb) erheblich.

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
	Gesundheit, öffentliche Verwaltung usw.).		
Sektoral (z. B. nationale Gesundheitsdatengesetze, besondere Kategorien nach DSGVO)	Domänenspezifische Datenklassifizierung und Verarbeitungsbeschränkungen.	Gesundheitswesen, Recht, bestimmte öffentliche Dienste.	Erzwingt oft Entscheidung 6 zu „Strikt"; Entscheidung 4 (Hardware) unberührt.
Nationales Souveränitätsmandat	Länderspezifisch (z. B. GAIA-X-Mitgliedschaft, Einstufung als souveräne Cloud). Oft am restriktivsten.	Öffentlicher Sektor, Verteidigung, kritische nationale Infrastruktur.	Entscheidung 6 auf „Strikt" oder „Air-Gapped" festgelegt; Entscheidung 7 typischerweise kundenbetrieben.

> Mehrere Regulierungen greifen häufig gleichzeitig. Eine Bank im Anwendungsbereich von DORA fällt in der Regel auch unter die DSGVO, unter sektorale nationale Aufsichtserwartungen und — bei EU-übergreifender Tätigkeit — unter NIS2. Die Architektur muss die restriktivste Vorgabe erfüllen, nicht den Durchschnitt.

Entscheidung 3 — Modellklasse

Open-Weight oder proprietär?

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Open-Weight (Llama, Mistral, Qwen, DeepSeek, Phi, Gemma usw.)	Kunde kontrolliert die Gewichte; kann feintunen; vollständig reproduzierbar; auditierbar. Lizenzbedingungen variieren (Llama-Lizenz vs. Apache 2.0 vs. andere — je Modell prüfen).	Souveränität oder Auditierbarkeit ist bindend. Kunde will Eigentum am Fine-Tuning. Kostenberechenbarkeit.	Entscheidung 4 (Hardware) auf das gewählte Modell dimensioniert; Entscheidung 6 (Souveränität) vollständig erreichbar.
Proprietär (Closed-Weight, herstellerlizenziert, z. B. GPT-4-Familie, Claude, Gemini)	Höhere Spitzenqualität bei bestimmten Aufgaben (heute). Anbieter kontrolliert Gewichte und Updates. Eingeschränktes Fine-Tuning. Auditierbarkeit begrenzt.	Qualitätsanforderungen von Open-Weight noch nicht erfüllt. Workload nicht im Regulierungsbereich.	Entscheidung 6 (Souveränität) stark eingeschränkt; Entscheidung 7 (Betrieb) selten kundenkontrolliert.

> Die Qualität von Open-Weight-Modellen hat den Abstand zu proprietären Modellen 2024–2026 deutlich verringert. Für die meisten Unternehmens-Workloads (RAG, Code-Unterstützung, strukturierte Generierung,

> Klassifizierung, Zusammenfassung) liefern Open-Weight-Modelle ab 70B produktionsreife Ergebnisse. Proprietäre Modelle führen weiterhin bei bestimmten Frontier-Aufgaben; je Anwendungsfall prüfen.

Entscheidung 4 — Hardware

Welche GPU-Klasse — dimensioniert für welche Workloads?

Option	Am besten für	Ungefähre Dimensionierung	Kompromisse
NVIDIA H100 / H200 (80 GB / 141 GB HBM)	Flaggschiff-Inferenz, Fine-Tuning großer Modelle, hoher Produktionsdurchsatz.	7B–70B Inferenz komfortabel; 70B–405B mit Multi-GPU-Sharding.	Höchste Kosten, längste Lieferzeit. Für Produktion im großen Maßstab gerechtfertigt.
NVIDIA A100 (40 GB / 80 GB)	Allzweck; kosteneffizient für 7B–70B Inferenz und kleines Fine-Tuning.	7B–13B komfortabel auf einer GPU; 70B mit Multi-GPU.	Ältere Generation; bessere Verfügbarkeit als H100. Gutes Preis-Leistungs-Verhältnis bei Inferenz.
NVIDIA L40S (48 GB)	Kosteneffiziente Inferenz, besonders für 7B–30B in Produktion.	7B–13B komfortabel; 70B erfordert sorgfältiges Sharding.	Kein NVLink — nicht für das Training großer Modelle geeignet. Hervorragende Inferenz-Kosten je Token.
NVIDIA Blackwell (B100/B200)	Größte Trainings- und Inferenz-Workloads (ab 2026).	405B+ Inferenz bei passender Clustergröße.	Neueste Generation; Verfügbarkeit begrenzt.
CPU-only / kleinere Beschleuniger (z. B. Intel Gaudi, AMD MI-Serie)	Kleine Modelle (≤7B), RAG-Retrieval, Embedding-Erzeugung, Batch-Inferenz.	Llama 7B Q4 auf CPU ist tragfähig; Embeddings auf CPU funktionieren gut.	Geringerer Durchsatz. Nützlich für Randworkloads und Souveränitäts-Sonderfälle (kein NVIDIA).

> Gemischte Hardware-Flotten sind üblich. L40S für kontinuierliche Inferenz plus H100 für Fine-Tuning-Spitzen ist ein häufiges Muster. Cozystack unterstützt heterogene GPU-Pools je Mandant.

Entscheidung 5 — Mandantenfähigkeit

Teilen sich mehrere Workloads, Kunden oder Geschäftsbereiche die Infrastruktur — und wie strikt sind sie isoliert?

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Single-Tenant	Am einfachsten. Eine Workload-Klasse pro Cluster. Höchste Kosten je Workload.	Einzelner kritischer Workload; keine kundenübergreifende Teilung nötig; einfachster Weg.	Entscheidung 6 (Souveränität) unkompliziert.
Multi-Tenant über Tenant-CRD	Gemeinsame Infrastruktur plus logische Isolation: Namespace, NetworkPolicy,	Mehrere Workloads, Geschäftsbereiche oder Kunden, die effizientes Ressourcen-	Am flexibelsten über alle anderen Entscheidungen.

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
(Cozystack-Muster)	Identität und Storage-Klasse je Mandant. Kosteneffizient.	Pooling mit nachweisbarer Isolation benötigen.	Für die meisten Fälle empfohlen.
Cluster-pro-Mandant	Strikte Isolation auf Cluster-Ebene. Höchster Betriebsaufwand. Höchste Kosten.	Feindliche Mandantenfähigkeit; souveräner Kunde mit Anspruch auf eigenen Cluster; regulatorische Erwartung physischer/virtueller Trennung.	Betrieblich aufwendig — Cluster-Lebenszyklus automatisieren (Cluster API / Cozystack-Provisionierung).

> Das Tenant-CRD ist das am besten skalierbare Muster, wenn Isolation logisch statt physisch sein darf. Es unterstützt Identitäts-, Netzwerk-, Storage- und Observability-Isolation je Mandant auf gemeinsamer Infrastruktur. Prüfen Sie die Erwartung der Regulierung, bevor Sie sich auf „Cluster-pro-Mandant" festlegen — es ist selten wörtlich vorgeschrieben.

Entscheidung 6 — Souveränität

Wie strikt muss das Deployment einschränken, wer den Workload einsehen, per Schlüssel kontrollieren, auditieren und ändern kann?

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Standard	Deployment in der Region (z. B. nur EU). Anbieter/Betreiber kann verwalten; kundenkontrolliert, wo sinnvoll.	Workload ohne striktes Souveränitätsmandat; kostensensitiv.	Mit den meisten anderen Optionen kompatibel.
Strikt	Kundenkontrollierte Verschlüsselungsschlüssel (BYOK / HYOK). Lieferantentransparenz bis zur zweiten Stufe. Audit-isolierte Umgebung. Zugriff des Anbieterpersonals protokolliert und zeitlich begrenzt.	Regulierte Workloads (DORA, NIS2). Souveräne Cloud-Kunden. Workloads mit nationalem Mandat.	Entscheidung 7 typischerweise kundenbetrieben oder hybrid mit striktem Vertrag.
Air-Gapped	Offline-Betrieb. Kein Call-home des Anbieters. Updates über signierte Bundles, out-of-band ausgeliefert.	Verteidigung, isolierte Industrieumgebungen, bestimmte Workloads der nationalen Sicherheit.	Zwingt Entscheidung 3 zu Open-Weight, Entscheidung 7 zu kundenbetrieben. Hoher Betriebsaufwand.

> „Kundenkontrollierte Schlüssel" müssen sich auf alle datentragenden Schichten erstrecken — Primärspeicher, Replikate, Backups, Observability-Daten, Modellgewichte im Ruhezustand und Trainingsdaten-Caches. Eine häufige Lücke sind verschlüsselte Produktionsdaten, während die Backups durch anbieterkontrollierte Schlüssel geschützt sind.

Entscheidung 7 — Betrieb

Wer betreibt die Plattform im Tagesgeschäft?

Option	Kompromisse	Wann wählen	Wechselwirkung mit anderen Entscheidungen
Kundenbetrieben	Kunde kontrolliert alles. Höchste Souveränität. Höchster interner Kompetenzbedarf.	Starkes Platform-Engineering-Team vorhanden; Air-Gap- oder strikte Souveränitätsanforderung; langfristiges Ziel der Eigenständigkeit.	Mit allen anderen Optionen kompatibel. Langsamster Weg zur ersten Produktion.
Anbieterbetrieben (von Aenix verwaltet)	Aenix betreibt die Plattform unter SLA. Kunde konzentriert sich auf Workloads. Schneller in Produktion.	Kein internes Plattformteam für KI-Infrastruktur; Produktionskapazität in 60–90 Tagen gewünscht; Souveränität über strikte vertragliche und technische Kontrollen erreichbar.	Schränkt die Souveränitätsoptionen von Entscheidung 6 ein, sofern Vertrag/Architektur nicht Kundenschlüssel und Audit-Export auf jeder Schicht garantieren.
Hybrid	Am häufigsten in regulierten Branchen: anbieterbetriebene Control Plane plus kundenbetriebene Data Plane (oder umgekehrt). Klare Schnittstelle vertraglich definiert.	Kunde will Kontrolle über Daten und Workloads behalten, aber den Betrieb beschleunigen. Häufig im DORA/NIS2-Kontext.	Am flexibelsten für regulierte Workloads. Erfordert vorab eine klare Schnittstellendefinition.

> Hybrid ist für regulierte Workloads oft die richtige Antwort. Der Kunde behält die Kontrolle über Daten, Identität, Verschlüsselungsschlüssel und Workload-Definitionen (auditierbar, exportierbar). Der Anbieter betreibt das Plattform-Substrat im vertraglichen Rahmen. Beide Seiten wirken bei der Vorfallbehandlung gemäß DORA Artikel 23 und NIS2 Artikel 23 mit.

Vier Referenzarchitekturen

Vier gängige Muster, vom kleinsten souveränen Fußabdruck bis zum maximal souveränen Air-Gapped-Deployment. Ordnen Sie Ihre sieben Entscheidungen dem nächstliegenden Muster zu.

Architektur 1 — Single-Tenant-Inferenzcluster

- Ein Kubernetes-Cluster, ein Kunde, eine Workload-Klasse (z. B. ein RAG-gestützter Chatbot).
- Komponenten: 4–8 Inferenzknoten (L40S- oder H100-GPU), 2 Control-Plane-Knoten, gemeinsamer Speicher (LINSTOR), Ingress (Cilium), Observability (VictoriaMetrics + VictoriaLogs).
- Einzelne Mandantengrenze auf Cluster-Ebene.

Fußabdruck: Kleinster tragfähiger Fußabdruck für souveräne Inferenz. Etwa 4–8 GPU-Knoten. Ein bis zwei Platform-Engineers in Teilzeit.

Architektur 2 — Multi-Tenant-Inferenzflotte (Tenant-CRD-Muster)

- Ein Cozystack-Cluster, mehrere Tenant-CRDs (jeweils = isolierter Namespace, NetworkPolicy, Storage-Klasse, Identität und Quota).
- 8–32 GPU-Knoten, vom Scheduler über die Mandanten hinweg geteilt.
- Observability-Ansichten je Mandant; mandantenübergreifende Nicht-Teilung durch NetworkPolicy erzwungen.

Fußabdruck: Skalierbare Multi-Kunden-Inferenz. Kosteneffizientes Pooling. Logische Isolation je Mandant nachweisbar.

Architektur 3 — Inferenz + Fine-Tuning + RAG

- Heterogener Cluster: H100/H200-Knoten für Fine-Tuning; L40S für kontinuierliche Inferenz; CPU-Knoten für RAG-Retrieval und Embedding.
- Vektor-DB (z. B. pgvector über den PostgreSQL-Operator oder Qdrant) für RAG.
- Objektspeicher (S3-kompatibel, MinIO oder ähnlich) für Trainingsdaten und Modell-Checkpoints.

Fußabdruck: Full-Stack-Muster für Organisationen, die sowohl Fine-Tuning als auch kontinuierliche Inferenz betreiben. Speicher plus Vektor-DB plus heterogene GPU.

Architektur 4 — Air-Gapped souveränes Deployment

- Isoliertes Netzwerk. Kein Internet-Egress. Updates über signiertes Bundle, importiert über eine Air-Gap-Bridge.
- Open-Weight-Modell (Llama / Qwen / Mistral). Eigenständige Registry (Harbor oder ähnlich). Selbst gehostete Observability.
- Kundenkontrollierte Schlüssel (HSM-gestützt). Alle Audit-Logs in ein kundenkontrolliertes SIEM exportiert.

Fußabdruck: Für Air-Gapped-souveräne Deployments. Hoher Betriebsaufwand; maximale Souveränität.

Sizing-Referenz

> Die folgenden Zahlen sind grobe Betriebsschätzungen zum Stand des Versionsdatums. Validieren Sie sie vor der Kapazitätsfestlegung anhand realer Werte. Aktualisieren Sie sie quartalsweise, da sich die Modell- und Hardwarelandschaft weiterentwickelt.

Sizing je Modell (Open-Weight, FP16-Basislinie; niedriger mit Quantisierung)

Modellfamilie	Größe	Min. GPU-Speicher (Einzelkarte)	Inferenz-TPS-Referenz	Anmerkungen
Llama-3-Familie	7B	16 GB (FP16) / 6 GB (Q4)	~80–120 TPS auf L40S	Einzelkarte auf den meisten modernen GPUs tragfähig
Llama-3-Familie	70B	140 GB (FP16) / 40 GB (Q4)	~25–40 TPS auf H100 (FP16)	Multi-GPU-Sharding für FP16; Q4 auf einer H100 tragfähig
Llama-3-Familie	405B	800 GB (FP16)	~15–20 TPS auf 8×H200	Produktionsmaßstab; Multi-Node üblich

Modellfamilie	Größe	Min. GPU-Speicher (Einzelkarte)	Inferenz-TPS-Referenz	Anmerkungen
Mistral-Familie	7B / Mixtral 8×7B	16 / 90 GB	ähnlich wie Llama-Äquivalente	MoE: weniger aktive Parameter als insgesamt
Qwen-Familie	7B / 14B / 72B / 110B	ähnlich wie Llama	ähnlich	Stark mehrsprachig; je Sprache prüfen
DeepSeek-Familie	7B / 67B / V3 (685B)	ähnlich wie Llama; V3 benötigt mindestens einen 8×H100-Cluster	variiert	Neuer; wettbewerbsfähig bei Qualität/Kosten
Phi-Familie (Microsoft)	3,8B / 14B	8 / 28 GB	~150+ TPS auf L40S	Starke Small-Model-Option für CPU-arme Szenarien
Gemma-Familie (Google)	2B / 9B / 27B	4 / 20 / 56 GB	ähnlich wie Llama 7B / 70B	Small-Model-Leistung, permissive Lizenz

Hardware-Kombinationen für typische Workloads

Workload-Profil	Empfohlene Hardware	Warum
Kontinuierliche Inferenz, 7B–13B, Einzelkunde	4× L40S-Knoten	Kosteneffizient, Einzelkarten-Serving, einfach linear skalierbar
Kontinuierliche Inferenz, 70B, Einzelkunde	2× H100 (NVLink) pro Replik, 2–4 Repliken	Gesharded Inferenz; starker Durchsatz
Multi-Tenant-Inferenzflotte, gemischte Modellgrößen	Heterogen: 8× L40S + 4× H100, je Workload gescheduled	Tenant-CRD weist je Workload-Klasse zu; hohe Pool-Auslastung
Fine-Tuning 7B–13B (LoRA / QLoRA)	1–2× H100 pro Fine-Tuning-Job	Speicher und Bandbreite für Gradienten-Updates
Fine-Tuning 70B (voll / partiell)	4–8× H100- / H200-Cluster	Multi-GPU und Multi-Node je nach Technik
Nur RAG-Retrieval + Embedding	CPU + kleine GPU (L40S / A100)	Embeddings auf CPU tragfähig; Retrieval ist I/O-gebunden
Air-Gapped souveränes Deployment	Open-Weight-Familie + On-Prem H100/L40S; vollständig selbst gehosteter Stack	Herstellerunabhängigkeit; maximale Souveränität

> Cozystack stellt die Plattformschicht bereit (Tenant-CRD-Mandantenfähigkeit, GPU-Scheduling, Observability, Identität, Storage). Ænix Platform bündelt dies mit einem Enterprise-SLA, GPU-Flottenbetrieb und Unterstützung für Fine-Tuning- und Inferenz-Workflow-Muster im großen Maßstab.

Häufige Fallstricke

- [ ] **„Souverän" mit „Air-Gapped" verwechseln.** Die meisten regulierten Workloads brauchen strikte Souveränität (kundenkontrollierte Schlüssel, Lieferantentransparenz, Audit-Isolation) — aber kein vollständiges Air-Gap. Der Betriebsaufwand eines Air-Gaps ist erheblich; wählen Sie ihn nur, wenn wirklich erforderlich.
- [ ] **Für Spitzenlast statt Dauerlast dimensionieren.** Produktionsinferenz läuft oft bei 30–60 % der Spitzenlast. Dimensionieren Sie für die Dauerlast mit Burst-Reserve (Autoscaling), nicht für die Spitze; andernfalls sind die GPU-Kosten je Token nicht wettbewerbsfähig.
- [ ] **Die Datenseite vergessen.** Vektor-DBs, Trainingsdatenspeicher und Observability-Daten brauchen alle Souveränität, Schlüsselkontrolle und Kapazitätsplanung. Häufig werden GPUs budgetiert und der Speicher vergessen.
- [ ] **Evaluierungen überspringen.** Die Auswahl von Open-Weight-Modellen sollte durch aufgabenspezifische Evaluierungen getrieben sein, nicht durch Schlagzeilen-Benchmarks. Bauen Sie eine kleine Evaluierungssuite für Ihre Workloads auf, bevor Sie sich auf eine Modellfamilie festlegen.
- [ ] **Abhängigkeit von einem einzelnen GPU-Anbieter.** NVIDIA dominiert 2026, doch eine Zweitquellen-Planung ist für Souveränität und Lieferkontinuität wichtig. Prüfen Sie zumindest, ob Open-Weight-Modelle auf alternativen Beschleunigern laufen (z. B. AMD MI300, Intel Gaudi).
- [ ] **Fine-Tuning-Betrieb unterschätzen.** Fine-Tuning ist ein Workflow, kein einmaliges Ereignis — Datenversionierung, Modell-Registry, Evaluierungs-Harness, Deployment-Pipeline. Planen Sie den Workflow, bevor Sie die Fine-Tuning-Hardware dimensionieren.

Arbeitsblatt: Entscheidungsübersicht

Halten Sie je Entscheidung Ihre gewählte Option und die Begründung fest. Ordnen Sie Ihr Profil anschließend der nächstliegenden Referenzarchitektur zu.

#	Entscheidung	Ihre Wahl	Warum?
1	Auslöser		
2	Regulierung		
3	Modellklasse		
4	Hardware		
5	Mandantenfähigkeit		
6	Souveränität		
7	Betrieb		

Nächstliegende Referenzarchitektur: Architektur 1 (Single-Tenant-Inferenz) · Architektur 2 (Multi-Tenant-Flotte) · Architektur 3 (Inferenz + Fine-Tuning + RAG) · Architektur 4 (Air-Gapped souverän).

Offene Fragen für die Teams aus Plattform, Sicherheit und Compliance: Listen Sie die Annahmen, von denen jede Entscheidung abhängt, die noch zu bestätigenden Erwartungen der Regulierung sowie die datenseitigen Kontrollen (Schlüssel, Backups, Observability), die noch nicht konzipiert sind.

Nächste Schritte

Option 1 — Selbst mit diesem Leitfaden planen. Arbeiten Sie ihn mit Ihrem Plattform- und KI-Team durch. Halten Sie die Entscheidungen im Arbeitsblatt fest. Bringen Sie offene Fragen in Ihre nächste Planungssitzung ein.

Option 2 — Souveräne KI-Architekturprüfung mit Aenix (1–2 Wochen). Wir prüfen Ihre Entscheidungen, validieren sie gegen die regulatorischen Anforderungen und erstellen eine Zielarchitektur-Skizze plus ein Sizing-Modell. Festpreis.

Option 3 — Phase 2: KI-Plattform-Aufbauprojekt. Mehrmonatiger Aufbau der Architektur: Cluster, GPU-Flotte, Mandantenfähigkeit, Identität, Observability, Model-Serving und Fine-Tuning-Workflow. Umfang nach Phase 1 festgelegt.

> Vereinbaren Sie ein Erstgespräch auf aenix.io. Aenix ist das Team hinter Cozystack (CNCF Project) und bietet Ænix Platform — unser kommerzielles, produktisiertes Angebot auf Basis von Cozystack.