XINITY

xinity.ai

VON GOOGLE GEMINI ZU XINITY

KI-PLATTFORM MIGRATIONS-WHITEPAPER SERIE 2026

LEGAL NOTICES

Xinity reminds you to carefully read through and completely understand all content in this section before you read or use this document. If you read or use this document, it is considered that you have identified and accepted all contents declared in this section.

1. This document is published by Xinity for informational purposes. The contents are intended for legal and compliant business activities. You shall not use or disclose all or part of the contents to any third party without written permission from Xinity.

2. This document may be subject to change without notice due to product upgrades, adjustment, and other reasons. Xinity reserves the right to modify the contents without notice.

3. This document is only intended for product and service reference. Xinity provides this document for current products and services with current functions, which may be subject to change.

4. All content including images, architecture design, page layout, and description text is owned by Xinity. You shall not use, modify, copy, or publish the content without written permission.

5. If you discover any errors or mistakes within this document, please contact Xinity directly.

THE AUTHORS

CORE AUTHORS

Alexander Zehetmaier (CEO & Co-Founder, Xinity)

TECHNICAL REVIEW

Jonas (CTO & Co-Founder, Xinity)

EDITING AND DESIGN

Xinity Marketing Team

TARGET AUDIENCE

Dieser Leitfaden richtet sich an Engineering-Teams, CTOs und IT-Entscheidungstraeger, die derzeit Googles Gemini KI-Dienste (Gemini 3.1 Pro, Gemini 2.5 Pro/Flash, Vertex AI, Google AI Studio) nutzen und KI-Workloads auf eine souveraene On-Premise-Infrastruktur migrieren muessen. Ob Sie Alternativen aufgrund von Datenresidenz-Bedenken innerhalb des GCP-Oekosystems evaluieren, regulatorischem Druck durch EU-Datensouveraenitaetsanforderungen ausgesetzt sind, oder strategisch die Abhaengigkeit von einem einzelnen Hyperscaler reduzieren moechten -- dieses Whitepaper liefert die technischen Zuordnungen und Migrationsprozesse.

CONTENTS

1. ENTERPRISE AI OHNE KOMPROMISSE: WARUM XINITY DIE BESSERE WAHL IST

Wenn Ihr Unternehmen KI-Workloads in der Produktion betreibt, bietet die Migration von Cloud-gehosteten KI-APIs zur On-Premise-Plattform von Xinity etwas, das kein Cloud-Anbieter liefern kann: vollstaendige architektonische Souveraenitaet ueber Ihre Daten, Modelle und Inferenz-Infrastruktur. Dies ist kein einfacher Anbieterwechsel -- es ist ein fundamentaler Wandel vom Mieten von KI-Kapazitaet zum Besitzen.

-- Architektonische Souveraenitaet statt Richtlinien-Versprechen

Cloud-KI-Anbieter bieten vertraglichen Datenschutz durch Nutzungsbedingungen und Auftragsverarbeitungsvertraege. Xinity liefert architektonische Souveraenitaet: Ihre Daten verlassen niemals Hardware, die Sie physisch besitzen und kontrollieren. Fuer regulierte Branchen -- Gesundheitswesen, Recht, Finanzdienstleistungen, Medien und Fertigung -- ist diese Unterscheidung nicht akademisch. Es ist der Unterschied zwischen Compliance-Risiko und Compliance-Sicherheit. Keine auslaendische Regierungsvorladung, keine Aenderung der Cloud-Anbieter-Richtlinien und keine geopolitische Verschiebung kann Daten beeinflussen, die ausschliesslich auf Ihren Raeumlichkeiten existieren.

-- Planbare Wirtschaftlichkeit im Enterprise-Massstab

Cloud-KI-Preise skalieren mit dem Verbrauch: Jeder API-Aufruf, jedes Token, jede GPU-Stunde wird gemessen und abgerechnet. Xinity's On-Premise-Modell wandelt variable OPEX in planbare CAPEX um. Kunden, die Xinity Runtime auf ASUS Ascent GX10 Servern einsetzen, berichten von ca. 80% Kostenersparnis gegenueber vergleichbarer Cloud-Kapazitaet. Im Massstab bedeutet das ca. 320 EUR/Jahr Stromkosten gegenueber 18.600 EUR/Jahr fuer vergleichbare Cloud-Rechenleistung.

-- Latenzfreie Inferenz fuer kritische Anwendungen

On-Premise-KI eliminiert Netzwerk-Roundtrips zu entfernten Cloud-Regionen. Fuer latenzsensitive Anwendungen -- Echtzeit-Dokumentenanalyse, Qualitaetskontrolle in der Produktion, klinische Entscheidungsunterstuetzung -- liefert lokale Inferenz konsistente Sub-Millisekunden-Antwortzeiten ohne Abhaengigkeit von Internetverbindung, Cloud-Region-Verfuegbarkeit oder grenzueberschreitenden Datentransfervorschriften.

-- Regulatorischer Rueckenwind beschleunigt die Adoption

Der EU Digital Networks Act (vorgeschlagen Januar 2026) mit Compliance-Fristen im August 2026, die 20 Milliarden EUR InvestAI-Foerderinitiative und aufkommende 'Buy European'-Beschaffungsregeln validieren die These der souveraenen KI-Infrastruktur. Organisationen, die jetzt auf On-Premise-KI migrieren, positionieren sich vor den Regulierungen statt spaeter hektisch reagieren zu muessen.

-- OpenAI-kompatible APIs -- migrieren ohne Neuentwicklung

Xinity Runtime stellt OpenAI-kompatible API-Endpunkte bereit. Das bedeutet: Ihr bestehender Anwendungscode, SDKs, Prompt-Bibliotheken und Orchestrierungsframeworks funktionieren mit minimalen Aenderungen weiter. Sie aendern die Base-URL und den API-Key; Ihre Anwendungen bemerken keinen Unterschied.

2. IHR GEMINI-STACK, NEU AUFGEBAUT AUF XINITY (ZUGEORDNET & BEREIT)

Dieser Abschnitt stellt ein Faehigkeiten-Mapping fuer die Migration von Googles Gemini-Oekosystem zu Xinitys On-Premise-Plattform bereit. Googles Gemini ist tief in GCP-Dienste integriert, daher erfordert die Migration sowohl API-Uebersetzung als auch architektonische Entkopplung.

Kern-Inferenz & Multimodale KI

Source Service	Xinity Equivalent	Migration Notes
Gemini 3.1 Pro (Neuestes Flaggschiff)	Xinity Runtime (Mistral Large 3 / Nemotron-Ultra)	OpenAI-kompatibler API-Endpunkt. Kontextfenster bis 128K Token. Fuer 1M+: Chunking + RAG-Pipeline.
Gemini 2.5 Pro (1M Token Kontext)	Xinity Runtime (Nemotron-Ultra / Qwen3.5 72B)	Komplexes Reasoning und Coding. Adaptive Denkfaehigkeiten. Lokale Inferenz zum Festpreis.
Gemini 2.5 Flash / Flash-Lite	Xinity Runtime (Qwen3.5 8B / Mistral Small 3)	Schnelle, kostenoptimierte Inferenz. Ideal fuer Klassifizierung und Zusammenfassung.

Vertex AI Plattform-Dienste

Source Service	Xinity Equivalent	Migration Notes
Vertex AI Model Garden	Xinity Model Registry	Kuratierter Open-Weight-Modellkatalog. Ein-Klick-Bereitstellung.
Vertex AI Pipelines	Xinity + Kubeflow / MLflow	On-Premise ML-Pipeline-Orchestrierung.
Vertex AI Feature Store	Xinity + Feast / Hopsworks	Selbst gehosteter Feature Store.
Vertex AI Workbench	Xinity + JupyterHub	On-Premise Notebook-Umgebung.

Embeddings & Suche

Source Service	Xinity Equivalent	Migration Notes
Gemini text-embedding-004	Xinity Runtime (BGE-M3, E5-Mistral)	Lokale Embedding-Generierung. Mehrsprachige Unterstuetzung.
Vertex AI Vector Search	On-Prem Vektor-DB (Qdrant / Milvus)	Selbst gehostete Aehnlichkeitssuche. Volle Datensouveraenitaet.

Plattform & Sicherheit

Source Service	Xinity Equivalent	Migration Notes
Google Cloud IAM	Xinity Admin Console (LDAP / SAML / OIDC)	On-Premise Identity-Integration. Keine Cloud-IAM-Abhaengigkeit.
Vertex AI Monitoring	Xinity + Prometheus / Grafana	Echtzeit-Modell-Performance-Tracking. Alle Metriken bleiben On-Premise.
Cloud Audit Logs	Xinity Audit-Modul	Vollstaendiger Inferenz-Audit-Trail. Compliance-Berichte fuer DSGVO.

3. MIGRATIONSPROZESS

3.1 Bestandsaufnahme & Discovery

Gemini & Vertex AI Nutzung auditieren

Exportieren Sie GCP-Abrechnungs- und Nutzungsberichte. Katalogisieren Sie alle Anwendungen, die Gemini-APIs aufrufen, und alle Vertex AI Pipelines.

GCP-Dienstabhaengigkeiten kartieren

Identifizieren Sie alle GCP-Dienstabhaengigkeiten (Cloud Storage, BigQuery, Pub/Sub) und planen Sie die Entkopplungsreihenfolge.

Workload-Souveraenitaet klassifizieren

Identifizieren Sie Workloads, die personenbezogene Daten unter DSGVO, Geschaeftsgeheimnisse oder regulierte Daten verarbeiten.

3.2 Infrastrukturplanung

Hardware-Dimensionierung

Dimensionieren Sie die Xinity-Bereitstellung basierend auf Ihrem Gemini-Nutzungsmuster.

GCP-Entkopplungsarchitektur

Entwerfen Sie die Zielarchitektur, die GCP-Dienste durch On-Premise-Aequivalente ersetzt.

API-Uebersetzungsschicht

Gemini-SDK-Aufrufe erfordern eine Uebersetzung zum OpenAI-Format. Xinity bietet ein leichtgewichtiges API-Gateway dafuer.

3.3 Pilot-Migration

Xinity Runtime bereitstellen

Installation auf Ihrer Hardware. API-Endpunkte und Modelle konfigurieren.

SDK-Migration

Anwendungscode von Googles Gemini SDK zu OpenAI-kompatiblen SDK-Aufrufen migrieren, die auf Xinity zeigen.

Parallele Validierung

Pilot-Workloads 2-4 Wochen parallel ausfuehren und Ergebnisqualitaet vergleichen.

3.4 Vollstaendige Migration

Phasenweise Workload-Migration

Souveraenitaetsblockierte Workloads zuerst, dann GCP-abhaengige Dienste, dann verbleibende API-Konsumenten.

GCP-Dienste ersetzen

On-Premise-Ersatz bereitstellen: Qdrant fuer Vector Search, Kubeflow fuer Pipelines, Feast fuer Feature Store.

GCP-Ressourcen dekommissionieren

Nach vollstaendiger Validierung GCP-Endpunkte beenden und Abrechnungskonten schliessen.

4. MIGRATIONS-WERKZEUGE & BESCHLEUNIGER

4.1 API-Uebersetzung & Kompatibilitaet

Gemini-zu-OpenAI API Gateway

Leichtgewichtiger Proxy, der Gemini-API-Anfragen ins OpenAI-Format uebersetzt.

SDK-Migrations-Toolkit

Automatisches Refactoring-Tool fuer die Konvertierung von Gemini SDK zu OpenAI SDK.

4.2 Observability & Betrieb

Xinity Dashboard

Vorkonfiguriertes Monitoring fuer alle On-Premise KI-Dienste.

Compliance & Audit-Modul

Compliance-Berichte fuer vollstaendige Datensouveraenitaet. DSGVO, ISO 27001.

5. NAECHSTE SCHRITTE: STARTEN SIE IHRE MIGRATION MIT XINITY

Die Migration von Google Gemini zu Xinity umfasst API-Uebersetzung und GCP-Dienst-Entkopplung, aber Xinitys OpenAI-kompatible Endpunkte machen den Uebergang systematisch.

So starten Sie:

1. Discovery-Gespraech vereinbaren -- Xinitys Solutions-Architekten kartieren Ihr gesamtes Gemini + GCP Oekosystem.

2. Proof of Concept anfordern -- Testen Sie die Gemini-zu-OpenAI API-Uebersetzung.

3. GCP-Entkopplung planen -- Sequenzielle Abloesung jedes GCP-Dienstes.

4. Go Live mit voller Souveraenitaet -- 100% Kontrolle ueber Daten und Infrastruktur.

Kontakt: Web: xinity.ai E-Mail: contact@xinity.ai Standort: Wien, Oesterreich