PRISKI · Lokale RAG-Architektur

Variante B – DSGVO-nativ, offline-fähig, im Portfolio-Kontext

Ein praxisgeprüftes Architektur-Whitepaper für datenschutzsensible Branchen: Kanzleien, Steuerberatungen, Versicherungen, Gesundheitswesen.

Autor: Fabian Kurzeja · Priski **Version:** 1.1 · April 2026 **Lesezeit:** ca. 14 Minuten **priski.de**

Priski · Whitepaper v1.1 · April 2026 · Interne Dokumentation, auszugsweise zitierbar mit Quellenangabe.

Warum dieses Whitepaper

Die meisten KI-Whitepaper im Mittelstand versprechen revolutionäre Produktivitätsgewinne und verlieren kein Wort darüber, was in der Praxis nicht funktioniert. Priski nimmt einen anderen Weg: Wir empfehlen keine Architektur, die wir nicht selbst unter Praxisbedingungen getestet haben.

Dieses Whitepaper beschreibt **Variante B** der Priski Ingestion Architecture – eine vollständig lokale, cloud-freie RAG-Architektur für datenschutzsensible Branchen. Es enthält:

1. Vier typische Einsatzszenarien aus der Praxis (Kanzlei, Ingenieurbüro, Anwaltskanzlei, Arztpraxis).
2. Eine vereinfachte 3-Schichten-Architektur ohne Airbyte, die auf einem normalen Büro-PC läuft.
3. Einen direkten Vergleich: Verteilte (A) vs. Lokale (B) Variante entlang sechs Kriterien.
4. Ehrliche Praxis-Learnings aus dem Priski-PoC-Test im April 2026 – inklusive der Grenzen lokaler LLMs und der konkreten Messwerte unseres Gold-Set-Tests (7,5 / 20).
5. Welche Hardware Sie brauchen – und wann ein Cloud-Hybrid die bessere Wahl ist.

Für wen?

Entscheider in Kanzleien, Steuerberatungen, Versicherungen, Arztpraxen und technologie-nahen Mittelstandsunternehmen, die vor der Frage stehen: *Können wir KI nutzen, ohne vertrauliche Daten in die Cloud zu geben – und zu welchem Preis?*

Dieses Whitepaper beschreibt **Variante B im Detail**. Für Teams mit anderen Datenschutz-Profilen empfehlen wir **Variante A (Cloud-Hybrid)** oder **Retrieval-Only** – mehr dazu in Kapitel 2.3.

Dieses Dokument ist Auszug aus der internen „Priski Ingestion Architecture" (Version 1.0, April 2026). Für den vollständigen Bericht inklusive Variante A (verteilte Cloud-Architektur), Embedding-Empfehlungen und ROI-Berechnung: discovery@priski.de.

---

1 · Variante B – Lokale Datenanalyse

Für Unternehmen, die keine Cloud-Anbindung benötigen oder wollen – etwa aus DSGVO-Gründen, wegen sensibler Daten oder schlicht weil alle relevanten Informationen lokal vorliegen – gibt es eine deutlich einfachere Architektur ohne Airbyte. Die Lösung läuft vollständig auf dem Rechner oder NAS des Kunden, benötigt keine Internetverbindung und kann auch externe Festplatten, USB-Sticks oder Netzlaufwerke direkt einlesen.

1.1 · Typische Einsatzszenarien

Vier Beispielbranchen zeigen, warum lokale Datenanalyse nicht akademisch ist, sondern konkreten Mehrwert liefert. In allen Fällen sind die Daten aus regulatorischen oder operativen Gründen nicht cloud-tauglich:

Branche	Datenquellen (lokal)	Typische Fragen an das System
Steuerkanzlei	Mandanten-Ordner (PDF, DOCX), externer NAS, USB-Archiv älterer Jahrgänge	„Welche Mandanten haben 2024 Umsätze über 500 T€ gemeldet?"
Ingenieurbüro	Technische Handbücher, Prüfberichte, Normdokumente auf lokalem Server	„Gibt es Prüfberichte zu Bauteil X mit Beanstandungen nach DIN 4102?"
Anwaltskanzlei	Vertragsarchiv auf externer Festplatte, gescannte Urteilssammlungen	„Welche Verträge enthalten Kündigungsklauseln ohne Abfindungsregelung?"
Arztpraxis / Klinik	Befundberichte, Entlassbriefe als PDF auf lokalem Server (komplett offline)	„Alle Patienten mit Diagnose X und Medikation Y im letzten Jahr"

1.2 · Vereinfachte Architektur (ohne Airbyte)

Da keine externen Systeme angebunden werden müssen, entfällt Airbyte vollständig. LlamaIndex liest die Dateien direkt über einen rekursiven Ordner-Scan ein – inklusive externer Festplatten, USB-Sticks und Netzlaufwerke (SMB/NFS). Als Vektordatenbank kommt ChromaDB zum Einsatz, die ohne separaten Datenbankserver läuft und direkt in die Anwendung eingebettet ist.

Schicht	Komponenten	Rolle
1 — Datenquellen (lokal)	Lokale Ordner (C:, D:), externe Festplatten, USB-Sticks, Netzlaufwerke (SMB/NFS)	Einmaliges Einbinden, rekursives Scannen. Kein Connector-Mapping.
2 — Verarbeitung	Unstructured.io (Parsing), LlamaIndex (SimpleDirectoryReader + Embedding-Pipeline)	PDF/DOCX/MD → Text-Chunks → Embedding-Vektoren. Open Source, keine Cloud-Kosten.
3 — Speicherung & Antwort	ChromaDB (embedded) + lokales oder EU-Cloud-LLM	Vektor-Suche in Millisekunden. Antwort-Generierung wahlweise offline (Ollama) oder via EU-API.

1.3 · Externe Festplatte anschließen – So funktioniert es

Der Kunde schließt die externe Festplatte per USB an den Rechner an, auf dem die Priski-Lösung läuft. Das System erkennt das neue Laufwerk und kann alle Dokumente darin in die Wissensbasis aufnehmen. Für Netzlaufwerke (NAS im Firmennetzwerk) wird lediglich der Netzwerkpfad einmalig eingetragen.

Ablauf in der Praxis:

1. Festplatte anschließen oder Netzpfad eintragen.
2. LlamaIndex scannt rekursiv alle unterstützten Formate (PDF, DOCX, XLSX, TXT, MD, CSV, EML).
3. Unstructured.io parst jede Datei in saubere Text-Chunks.
4. Embeddings werden lokal (Ollama) oder über EU-API berechnet und in ChromaDB gespeichert.
5. Wissensbasis ist sofort durchsuchbar – Antworten in wenigen Sekunden.

1.4 · Vergleich: Lokale vs. Verteilte Variante

Die lokale Variante ist nicht besser oder schlechter als die verteilte – sie adressiert ein anderes Problem. Entscheidend ist, welcher Typ Kunde vor uns sitzt:

Kriterium	Variante A – Verteilt (mit Airbyte)	Variante B – Lokal (ohne Airbyte)
Datenquellen	350+ Konnektoren: ERP, CRM, Cloud, Datenbanken, APIs	Lokale Ordner, externe HDD/SSD, NAS, USB
Internetzugang	Erforderlich (Airbyte-Konnektoren + LLM-API)	Nicht notwendig (100 % Offline-Betrieb möglich)
Infrastruktur	Cloud oder dedizierter Server (AWS / Azure)	Vorhandener PC oder NAS genügt (min. 8 GB RAM)
Laufende Betriebskosten	80–180 €/Monat (ohne Priski-Retainer)	0–15 €/Monat (nur optionale LLM-API-Kosten)
Einrichtungsaufwand	3–10 Tage (abhängig von Anzahl Quellsysteme)	1–2 Tage (einfaches Setup, kein Connector-Mapping)
Empfohlen für	KMU mit Daten in mehreren Systemen (ERP, CRM, E-Mail, SharePoint)	Kanzleien, Praxen, Archiv-Projekte mit sensiblen / lokalen Daten

> **Priski-Entscheidungs-Regel:** Wenn ≥ 3 verteilte Quellsysteme angebunden werden müssen → Variante A. Wenn alle kritischen Daten auf lokalen Servern / Festplatten liegen und DSGVO-Härte oberste Priorität hat → Variante B. In ca. 70 % der Mittelstandsfälle ist eine **Hybrid-Lösung** (Embeddings lokal, LLM in EU-Cloud) der pragmatische Sweet Spot – diese Variante beschreibt Kapitel 2.3 ausführlich.

---

2 · Praxis-Learnings aus eigenem PoC-Test (April 2026)

Priski hat die vollständig lokale Variante (Ollama + LlamaIndex + pgvector) im April 2026 auf einem Konsumenten-Setup getestet – mit realen deutschsprachigen Dokumenten (Unternehmenskonzept, technische Setup-Dokumente, Whitepaper, ca. 80 Dateien in MD/DOCX/PDF/XLSX). Die ehrlichen Erkenntnisse:

2.1 · Was funktioniert

- **Ingestion und Retrieval laufen zuverlässig.** Mit `mxbai-embed-large` (1.024 Dimensionen) oder `jina-embeddings-v2-base-de` werden deutsche Dokumente sauber vektorisiert und passende Chunks korrekt gefunden. Retrieval-Scores im Gold-Set-Test lagen durchgängig zwischen **0,66 und 0,83** – die richtigen Dokumente werden zuverlässig identifiziert.
- **Parsing mit Unstructured.io funktioniert zuverlässig** – PDFs, DOCX und Markdown werden in saubere Text-Chunks überführt.
- **Der Ingestion-Teil ist produktionsreif** – die Pipeline ist stabil und reproduzierbar. Wer eine lokale Semantik-Suche über Akten, Berichte oder Archive braucht, bekommt sie.

2.2 · Was (noch) nicht reicht

- **Antwortqualität kleiner lokaler LLMs (Llama 3.1 8B) ist auf Deutsch schwach** – oberflächlich, halluziniert, wechselt zurück auf Englisch.

- **Größere Modelle (Gemma 4 27B) liefern brauchbare Qualität** – benötigen aber auf Konsumenten-Hardware mehrere Minuten pro Antwort. Auf typischer Mittelstand-Hardware (Windows-Büro-PCs ohne GPU) ist das nicht produktionstauglich.
- **Für akzeptable Performance eines lokalen LLMs braucht es eine dedizierte GPU** (RTX 4090 oder äquivalent) – also Hardware-Investitionen von 5.000–15.000 € einmalig.
- **Konkreter Messwert:** In unserem Gold-Set-Test (20 vorbereitete Prüf-Fragen gegen das eigene Korpus, Zielhitrate ≥ 14/20) erreichte die 8B-Konfiguration **7,5 / 20 = 37,5 %**. Detailauswertung folgt in Kapitel 2.3, Block D.

2.3 · Grenzen lokaler Setups – wann Cloud-Hybrid oder Retrieval-Only die bessere Wahl ist

Block A — Die ehrliche Antwort auf „Für wen ist Variante B?" Variante B ist technisch das sauberste Konstrukt: keine Cloud-Abhängigkeit, kein Drittanbieter im Datenfluss, keine API-Limits. Aber gerade weil sie keine Abkürzungen kennt, ist sie nicht für jeden die richtige Wahl. Hardware-Investment, Modell-Größe und Use-Case-Anspruch müssen zueinander passen. Wer Variante B aus Reflex wählt, nur weil sie „am sichersten" klingt, kann mit Konsumenten-Hardware schnell auf Antwortqualitäten landen, die im Tagesgeschäft enttäuschen. Diese Sektion beschreibt, **wann Variante B trägt – und wann eine andere Variante das bessere Werkzeug ist.**

Block B — Drei Varianten, ein Entscheidungs-Schema Priski bietet seit April 2026 drei Pilot-Varianten an. Welche zum Kunden passt, entscheidet der Discovery-Call gemeinsam mit Datenschutz-Anspruch und Hardware-Realität:

Variante	Setup (Kurzform)	Zielkunde	Preisanker (Pilot, 4 Wochen)
B-Enterprise *(dieses Whitepaper)*	Lokaler LLM ≥ 32B, dedizierter Server (GPU-Klasse RTX 4090 oder besser)	Behörden, Kanzleien mit On-Prem-Pflicht, Berufsgeheimnis-Kontext	**6.900 – 9.900 €** *(Hardware nicht inkl.)*
A-Cloud-Hybrid	Lokale Ingestion / Embedding / Vector-Store, Generator via EU-Cloud-LLM (z. B. GPT-4o-mini via EU-DPA, Claude Haiku in EU-Region)	Mittelstand, Steuerberater, KMU mit pragmatischer AV-Vertrags-Bereitschaft	**3.900 – 4.900 €** + ca. 5–20 €/Monat Cloud-LLM
Retrieval-Only	Lokale Semantik-Suche, Top-5-Quellen mit Snippet, keine LLM-Antwort-Generierung	Archiv-Sucher, QM-Abteilungen, KI-skeptische Häuser, Behörden ohne KI-Mandat	**2.900 – 3.900 €**

Die Varianten sind **austauschbar**: Wer mit Retrieval-Only einsteigt, kann später auf A-Cloud-Hybrid hochrüsten, sobald die Antwort-Generierung gewünscht ist. Wer mit B-Enterprise startet, kann einzelne Workloads in den Cloud-Hybrid verlagern, sobald die Workload-Mischung sich verschiebt. Der Einstieg sollte aber bewusst am tatsächlichen Datenschutz-Anspruch und am vorhandenen Hardware-Budget gewählt werden – nicht am Wunschbild.

Block C — Wann ist Variante B die richtige Wahl? Vier harte Kriterien sprechen für Variante B-Enterprise als Erst-Wahl:

1. **Berufsgeheimnis oder Verschlusssachen-Klassifizierung** – etwa § 203 StGB (Berufsgeheimnis), StBerG (Steuerberatungsgesetz), RAO (Anwaltsordnung) oder vergleichbare Compliance-Mandate, bei denen ein AV-Vertrag mit einem Cloud-Anbieter nicht ausreicht oder politisch nicht durchsetzbar ist.
2. **Kein AV-Vertrag mit Cloud-Anbieter möglich oder gewollt** – etwa weil das Haus konsequent auf On-Prem-Strategie setzt oder weil Mandanten-Verträge Cloud-Verarbeitung explizit ausschließen.
3. **Hardware-Budget ≥ 8.000 €** für einen dedizierten GPU-Server (oder eine entsprechende Workstation), der die LLM-Inferenz im 30B–70B-Bereich tragen kann.
4. **Datenvolumen > 50 GB** mit kontinuierlichem Wachstum. Hier lohnt sich dauerhafte On-Prem-Infrastruktur, weil Daten-Upload-Aufwände in Cloud-Hybrid-Szenarien sonst zur Reibung werden.

Wer drei oder mehr dieser Kriterien voll erfüllt, sollte Variante B ernsthaft prüfen. Wer keines oder nur eines erfüllt, fährt mit **A-Cloud-Hybrid** oder **Retrieval-Only** in der Regel besser – günstiger, schneller produktiv, mit weniger Hardware-Bindung.

Block D — Der Gold-Set-Test: Was uns 7,5 / 20 gelehrt hat Im April 2026 haben wir das vollständige Variante-B-Setup auf Konsumenten-Hardware (RTX 4070 / 8 GB-Klasse, Ollama mit `llama3.1:8b` als Generator) getestet. Gemessen wurde gegen ein **20-Fragen-Gold-Set** auf das eigene Unternehmens-Korpus (~ 80 Dateien aus Konzepten, Setup-Dokumenten und Whitepapern). Zielhitrate: 14 / 20 (70 %).

Ergebnis: 7,5 / 20 = 37,5 %. Klassifizierung: 4 PASS · 7 PARTIAL · 9 FAIL.

Das ist kein einheitliches Versagen. Die Retrieval-Ebene hat in allen 20 Fragen die richtigen Dokumente in die Top-5 geholt (Score-Range 0,66–0,83). **Die Schwäche liegt beim Generator-Modell, nicht beim Retrieval.** Das 8B-Modell hat in mehreren Fällen die korrekt vorgelegten Quellen nicht genutzt – entweder weil Zahlen aus Tabellen nicht extrahiert wurden, weil Chunk-Größen Listenstrukturen zerrissen, oder weil das Modell selbstbewusst halluzinierte (etwa „der Kalkulator kann wahrscheinlich kostenlos heruntergeladen werden", obwohl das Gegenteil im Korpus stand).

Schluss-Haltung: **Wir empfehlen keine Architektur, die wir nicht selbst unter Praxisbedingungen getestet haben – und deshalb empfehlen wir Variante B nicht als Default, sondern zielgerichtet.** Für Mittelstands-Einsätze ohne harte On-Prem-Pflicht ist A-Cloud-Hybrid qualitativ und wirtschaftlich überlegen. Für Häuser mit echtem Berufsgeheimnis-Anspruch und ausreichendem Hardware-Budget bleibt Variante B die saubere Wahl – aber dann mit Modellen ab 32B, nicht mit Konsumenten-Klasse.

> **Was sich bis Q4 2026 ändern könnte:** Erscheint ein neues 7–8B-Open-Source-Modell, das deutsche QA signifikant besser beherrscht (Kandidaten: nächste qwen-Generation, Mistral-Next, DeepSeek), oder fällt die 24-GB-GPU-Klasse unter 1.000 €, wiederholen wir den Gold-Set-Test und aktualisieren diese Empfehlung. Review-Turnus: 6 Monate.

Block E — Was Sie im Discovery-Call entscheiden Sie müssen die Varianten-Frage **nicht alleine** entscheiden. In einem 30-minütigen Discovery-Call (kostenlos) klären wir gemeinsam, welche Variante zu Ihrem Setup passt. Mitbringen müssen Sie:

- eine kurze Beschreibung Ihres Use Cases (welche Frage soll die KI beantworten?),
- eine grobe Schätzung des Datenvolumens (GB-Bereich, Dateitypen),
- Ihren Datenschutz-Anspruch (Berufsgeheimnis-Mandat? AV-Vertrag möglich? On-Prem-Pflicht?).

Alles andere – Hardware-Empfehlung, Preisanker, Pilot-Zuschnitt – erarbeiten wir im Gespräch. Sie verlassen den Call mit einer klaren Empfehlung, nicht mit einem Vertriebs-Skript.

> **Fazit:** Genau diese ehrliche Trade-off-Analyse ist das, was Priski von reinen Folien-Beratern unterscheidet. Unser Selbst-Pilot-Test 2026 zeigt ehrlich, wo Variante B liefert

und wo sie an Grenzen stößt – und deshalb empfehlen wir sie nur dort, wo sie auch unter Praxisbedingungen überzeugt.

Wie geht es weiter?

Dieses Whitepaper ist der Einstieg – nicht das Ende. Wenn Sie in Ihrer Organisation konkret prüfen wollen, ob eine lokale, hybride oder retrieval-basierte RAG-Architektur für Ihre Daten funktioniert, bietet Priski drei strukturierte Einstiegs-Formate:

Das Priski-Angebot (auf einen Blick)

Im Discovery-Call entscheiden wir gemeinsam, welche Variante zu Ihrem Setup passt.

Format	Inhalt	Dauer	Investition
1 · Discovery-Call	30 Min. Telefon / Video. Wir schauen, ob Ihr Use Case zu unserem Angebot passt – ehrlich, ohne Pitch. Sie verlassen den Call mit einer Varianten-Empfehlung.	30 Min.	**kostenlos**
2 · Assessment *(optional)*	Priski analysiert vor Ort (oder remote) Ihre Datenlage, identifiziert Quick-Wins, dokumentiert Quellsysteme und liefert einen Architektur-Vorschlag inkl. Hardware-Empfehlung.	1 Tag	**1.800 €**
3 · Pilot (Festpreis)	In 4 Wochen: lauffähiger RAG-Assistent auf Ihren Daten. Inklusive Schulung, Übergabe-Doku, 30 Tage Hypercare. Preis je nach Variante:	4 Wochen	**siehe unten**

Pilot-Preise nach Variante (Festpreis, 4 Wochen):

- **Variante A · Cloud-Hybrid** – 3.900 – 4.900 €
- **Variante B · Enterprise (lokal, ≥ 32B-LLM)** – 6.900 – 9.900 € *(Hardware nicht inkl.)*
- **Variante Retrieval-Only** – 2.900 – 3.900 €

Welche Variante zu Ihrem Datenschutz-Anspruch und Ihrem Hardware-Budget passt, klären wir im Discovery-Call. Sie müssen nicht vorab entscheiden.

Eat your own dog food

Priski nutzt die beschriebene lokale RAG-Architektur selbst – für das interne Wissensmanagement (Konzepte, Setup-Dokumente, Kundenhistorie). **Jeder Empfehlung, die wir aussprechen, ist ein eigenes Build vorausgegangen.** Das ist unser Qualitäts-Versprechen, nicht unser Marketing-Satz. Auch der Gold-Set-Test in Kapitel 2.3 ist Teil dieses Prinzips: Wir messen unser eigenes System ehrlich – und ziehen die Konsequenzen.

Kontakt aufnehmen

Fabian Kurzeja — Gründer, Priski

- E-Mail: discovery@priski.de
- Web: priski.de
- LinkedIn: linkedin.com/in/fabian-kurzeja
