Solution Brief - Production-Grade RAG Agent

Problem

Business knowledge is split across docs, databases, and APIs.
Generic chatbots guess; enterprises need grounded answers.

RAG approach

1) Ingest data (PDF/DOCX/XLSX/CSV/DB/API)
2) Embed and store in vector DB (Milvus or in-memory)
3) Retrieve top context and answer only from that context
4) Refuse if context is insufficient

Key capabilities

- FastAPI backend with /ingest and /query endpoints
- Multi-tenant isolation + API key roles
- Audit logs and ingestion tracking
- Configurable embeddings (OpenAI/Gemini/local)
- Guardrails: no context -> no answer

Deliverables

- Source code + Docker stack (Standard/Premium)
- Configuration guide and deployment notes
- Sample requests, responses, and testing checklist

Deployment options

- Local Docker, cloud VM, or on-prem
- Milvus + Postgres supported in Premium