Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Production RAG Systems for Engineering Teams

Class Duration

14 hours of live training delivered over 2-3 days to accommodate your scheduling needs.

Student Prerequisites

- Professional software development experience in Python or TypeScript
- Basic familiarity with databases and REST APIs

Target Audience

Software engineers and ML engineers building internal knowledge bases, document Q&A systems, or AI-powered search features on top of organizational data. Relevant for teams that need to go beyond simple vector search demos and deploy reliable RAG systems that perform well on real enterprise data.

Description

This course covers production-grade RAG architecture from ingestion pipeline to deployed application. We go well beyond the naive chunk-embed-retrieve-generate pattern to cover the techniques that actually matter for production quality: advanced chunking (semantic, hierarchical, late chunking), hybrid search, cross-encoder reranking, query transformation, multi-turn conversation with retrieval, evaluation harnesses with golden datasets, and deployment patterns. Labs build a complete, evaluated RAG system on realistic enterprise-style document and data sources.

Learning Outcomes

- Design a production RAG ingestion pipeline with appropriate chunking and metadata strategies.
- Implement hybrid search combining dense retrieval and BM25 with result fusion.
- Apply cross-encoder reranking to improve retrieval precision.

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

- Use query transformation techniques (HyDE, query expansion, step-back) to improve recall.
- Build a RAG evaluation harness with a golden dataset, measuring RAGAS-style metrics.
- Identify and remediate common RAG failure modes from an evaluation run.
- Deploy a RAG API service with appropriate caching, cost controls, and observability.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Python 3.12+, Docker, API keys for an embedding model and a frontier LLM, and Git.

Training Topics

RAG Architecture for Production

- Ingestion pipeline, retrieval pipeline, and generation layer
- Where naive RAG fails in production
- Design decisions that determine RAG quality

Advanced Chunking

- Semantic chunking with embedding similarity
- Hierarchical chunking: parent/child retrieval
- Late chunking for long-context models
- Contextual Retrieval: prepending an LLM-generated context summary to each chunk before embedding (Anthropic, 2024–2026 standard)
- Late chunking vs. Contextual Retrieval: efficiency vs. relevance tradeoff
- Metadata extraction and storage

Vector Stores and Indexing

- pgvector, Qdrant, Weaviate, and Pinecone comparison
- Index configuration: HNSW vs. IVF tradeoffs
- Incremental and batch indexing pipelines

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

- Multi-tenancy and access control

Hybrid Search

- BM25 keyword search alongside dense retrieval
- Reciprocal Rank Fusion and score normalization
- Sparse-dense hybrid index options

Reranking and Query Transformation

- Cross-encoder rerankers: quality vs. latency
- HyDE (Hypothetical Document Embeddings)
- Query expansion and step-back prompting
- Multi-query retrieval for robustness

Multi-Turn RAG

- Conversation-aware retrieval with history
- Context window management over turns
- Follow-up query resolution

RAG Evaluation

- Evaluation dimensions: faithfulness, answer relevance, context recall
- RAGAS-style metrics and tooling
- Building a golden dataset
- Running evals in CI to catch regressions

Deployment and Operations

- RAG API service architecture
- Caching: embedding cache and response cache
- Cost attribution per request
- Monitoring retrieval quality in production

