+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 04/13/2026

Sign up

GraphRAG Training: RAG On-Premise with Neo4j

5 days (35 hours)

Overview

Our GraphRAG Training: On-Premise RAG with Neo4j will enable you to design, deploy, and optimize AI systems capable of querying your confidential data with surgical precision.

You will learn to master the entire technical stack, from high-performance serving with vLLM to vector indexing with Qdrant, to best meet the requirements of your on-premises infrastructure.

You will discover the fundamental concepts that make up modern RAG: ingestion via industrial OCR, semantic chunking, and re-ranking strategies.

We will work with management interfaces and APIs to quickly develop your first pipeline capable of extracting precise knowledge from heterogeneous document corpora.

Finally, you will master the advanced concepts essential for moving to production. Through hands-on practice, we will explore how to overcome the limitations of vector search by leveraging a Knowledge Graph to implement a GraphRAG strategy with Neo4j. You will also learn to manage the reliability of your responses and eliminate hallucinations using the RAGAS evaluation framework.

In short: move from prototype to industrial expertise by mastering the entire RAG stack
!

Objectives

- Deploy and optimize sovereign LLMs in a local environment

- Design complex ingestion pipelines including OCR and normalization
- Architect scalable and high-performance vector stores with Qdrant
- Implement the GraphRAG approach to address complex reasoning needs
- Measure and ensure the reliability of responses using scientific evaluation protocols

Target Audience

- Data Scientists
- Data Engineers
- Machine Learning Engineers
- Solution Architects
- Cloud Engineers
- CIO

Requirements

- Proficiency in Python
- Basic knowledge of Docker and Linux environments
- Theoretical knowledge of NLP or Machine Learning

GraphRAG Training Program: On-Premise RAG with Neo4j

[Day 1 - Morning]

Infrastructure and Local LLM Serving

- Secure architecture: network isolation, GPU quota management, and vRAM optimization
- High-Performance Serving with vLLM: Continuous Batching and PagedAttention
- Quantization techniques (AWQ, GPTQ, GGUF) to maximize hardware throughput
- Advanced inference parameter management: Temperature, Top-P, and context window
- Hands-on workshop: Deploying a local LLM and comparative performance benchmarking.

[Day 1 - Afternoon]

Ingestion and Preprocessing of Critical Data

- Multi-source extraction (complex PDFs, tables, images) and industrial OCR
- Cleaning and normalization: handling encodings and removing document noise
- Advanced chunking strategies: semantic, structural, and sliding window with overlap
- Automatic data integrity validation and source metadata management
- Hands-on workshop: Building a robust ingestion pipeline for heterogeneous documents.

Vector Store architecture with Qdrant

- Deploying Qdrant: Sharding, replication, and High Availability (HA) strategies
- Tuning HNSW indexes: balancing search speed, memory, and accuracy
- Persistence management, snapshots, and disaster recovery mechanisms
- Modeling the Payload schema for efficient large-scale filtering

[Day 2 - Morning]

Optimizing Retrieval and Embeddings

- Fine-tuning embedding models (Bi-Encoders) on domain-specific datasets
- Hybrid search: merging dense (vector) and sparse (BM25) scores
- Advanced metadata filtering techniques (Boolean, numerical, full-text)
- Score normalization for hybrid fusion (Reciprocal Rank Fusion)
- Hands-on workshop: Optimizing recall on a real-world domain corpus.

[Day 2 - Afternoon] Reranking

and Post-processing

- Integration of re-ranking models (cross-encoders such as BGE or ColBERTv2)
- Context window management and semantic compression of prompts
- Adaptive query rewriting to resolve ambiguities
- Detection and elimination of semantic duplicates in results

Knowledge graph modeling

- Limitations of semantic similarity: why vectors are no longer sufficient for global relationships
- Modeling strategic entities, relationships, and events with Neo4j
- Installing and securing the graph database in an isolated local environment
- Designing extensible graph schemas compatible with AI reasoning

[Day 3 - Morning]

Graph Extraction and Construction

- Automated extraction of complex entities and relationships using local LLMs
- Mastering the Cypher language for querying and manipulating knowledge
- Deduplication and entity resolution for graph consistency
- Validation of the graph structure against the defined business schema
- Hands-on workshop: Automatic population of a Knowledge Graph from the document stream.

[Day 3 - Afternoon] GraphRAG

Hybrid Reasoning

- Implementing GraphRAG: Merging Vector and Structural Search
- Navigating the graph to enrich the context
- JSON Schema Management and Function Calling for Structured Responses
- Pathfinding and community detection algorithms for global summarization
- Hands-on workshop: Developing a response engine that links facts across distant documents.

Orchestration with LlamaIndex and LangChain

- Design of orchestration patterns: Query Router and Sub-query Engine
- State management and long-term memory
- Chain observability: step traceability and fine-grained management of execution logs
- Breaking down complex queries into parallelizable subtasks

[Day 4 - Morning]

Agentic RAG and Autonomy

- Designing agents capable of planning and invoking external tools
- Reflection loops to minimize errors
- Automated generation of multi-document reports structured according to templates
- Handling information conflicts between contradictory sources
- Hands-on workshop: Creating an autonomous agent.

[Day 4 - Afternoon] Scientific

Evaluation

- RAGAS Framework: Faithfulness, Answer Relevance, and Context Precision metrics
- Performance audit: isolating search errors from synthesis errors
- Creating reproducible testbeds with "Gold Datasets"
- Calculating cost and latency per query for economic optimization

Interpretable AI and Trust

- Clear citation mechanisms and real-time source verification
- Display of confidence scores and chains of thought
- Human-in-the-loop interface for continuous improvement
- Explanability of results: analysis of the importance of retrieved documents

[Day 5 - Morning]

Performance and Semantic Caching

- Implementing Semantic Caching to reduce response times and GPU usage
- Optimization of "Time To First Token" latency and overall throughput
- Pre-computation strategies and query parallelization
- Queue management to handle load spikes

[Day 5 - Afternoon] LLMOps

and Lifecycle

- Hardware resource monitoring and semantic drift detection
- CI/CD for AI: automated semantic non-regression testing
- Model version management and reindexing without downtime
- Automation of feedback loops and collection of business usage metrics

Security and Production Playbook

- Protection against prompt injections and implementation of output guardrails
- Granular access control and on-premises encryption
- Development of the production playbook: backup and restore procedures
- GDPR compliance in a sovereign environment and physical data isolation
- Hands-on Workshop: Comprehensive Audit, System Hardening, and Security Protocol Validation.

Target Audience

This training is intended for both individuals and companies, large or small, seeking to train their teams in new advanced IT technologies or to acquire specific business knowledge or modern methods.

Assessment upon enrollment

The pre-training assessment complies with Qualiopi quality standards. Upon final registration, the learner receives a self-assessment questionnaire that allows us to evaluate their estimated proficiency in various types of technologies, as well as their expectations and personal goals regarding the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could pose challenges for monitoring and ensuring the smooth running of the training session.

Teaching Methods

Practical Course: 60% Practical, 40% Theory. Training materials distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the trainer, supported by examples and reflection sessions, and group work.

Assessment

At the end of the session, a multiple-choice questionnaire is used to verify that the skills have been properly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training program.

Training organization registered under number 11 75 54743 75. This registration does not constitute state accreditation.
© Ambient IT 2015-2026. All rights reserved. Paris, France - Switzerland - Belgium - Luxembourg