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

Mis à jour le 09/04/2026

S'inscrire

Formation GraphRAG : RAG On-Premise avec Neo4j

5 jours (35 heures)

Présentation

Notre formation GraphRAG : RAG on-premise avec Neoj4 vous permettra de concevoir, déployer et optimiser des systèmes d'IA capables d'interroger vos données confidentielles avec une précision chirurgicale.

Vous apprendrez à maîtriser l'ensemble de la stack technique, du serving haute performance avec vLLM à l'indexation vectorielle sous Qdrant, afin de coller au mieux aux exigences de votre infrastructure locale.

Vous découvrirez les concepts fondamentaux qui composent le RAG moderne : l'ingestion par OCR industriel, le chunking sémantique, ainsi que les stratégies de Re-ranking.

Nous manipulerons les interfaces de gestion et les APIs afin de développer très rapidement votre premier pipeline capable d'extraire des connaissances précises depuis des corpus documentaires hétérogènes.

De plus, vous apprendrez les concepts avancés indispensables à la mise en production. Nous verrons, à travers la mise en pratique, comment dépasser les limites du vectoriel grâce au Knowledge Graph (GraphRAG) avec Neo4j, et comment piloter la fiabilité de vos modèles via le framework d'évaluation RAGAS.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Déployer et optimiser des LLMs souverains en environnement local.
- Concevoir des pipelines d'ingestion complexes incluant l'OCR et la normalisation.
- Architecturer des Vector Stores scalables et performants avec Qdrant.
- Implémenter l'approche GraphRAG pour résoudre les besoins de raisonnement complexe.
- Mesurer et garantir la fiabilité des réponses via des protocoles d'évaluation scientifique.

Public visé

- Data Scientists
- Data Engineers
- Machine Learning Engineers
- Architectes Solutions
- Cloud Engineers
- DSI

Pré-requis

- Maîtrise du langage Python
- Bases en Docker et environnements Linux
- Notions théoriques en NLP ou Machine Learning

Programme de la formation GraphRAG : RAG On-Premise avec Neo4j

[Jour 1 - Matin]

Infrastructure et Serving LLM local

- Architecture souveraine : isolation réseau, gestion des quotas GPU et optimisation vRAM
- Serving haute performance avec vLLM : Continuous Batching et PagedAttention
- Techniques de quantification (AWQ, GPTQ, GGUF) pour maximiser le débit hardware
- Gestion des paramètres d'inférence avancés : Temperature, Top-P, et context window
- Atelier pratique : Déploiement d'un LLM local et benchmark de performance comparative.

[Jour 1 - Après-midi]

Ingestion et Pré-traitement de données critiques

- Extraction multi-sources (PDF complexes, tableaux, images) et OCR industriel
- Nettoyage et normalisation : gestion des encodages et suppression du bruit documentaire
- Stratégies de Chunking avancé : sémantique, structurel et fenêtrage glissant avec recouvrement
- Validation automatique de l'intégrité des données et gestion des métadonnées sources
- Atelier pratique : Construction d'un pipeline d'ingestion robuste pour documents hétérogènes.

Architecture de Vector Store avec Qdrant

- Déploiement de Qdrant : Sharding, réplication et stratégies de Haute Disponibilité (HA)
- Tuning des index HNSW : équilibrage entre vitesse de recherche, mémoire et précision
- Gestion de la persistance, des snapshots et mécanismes de récupération après sinistre
- Modélisation du schéma de Payload pour un filtrage efficace à grande échelle

[Jour 2 - Matin]

Optimisation du Retrieval et Embeddings

- Fine-tuning de modèles d'embeddings (Bi-Encoders) sur des jeux de données métier
- Recherche hybride : fusion des scores denses (vecteurs) et parcimonieuse (BM25)
- Techniques de filtrage avancé par métadonnées (booléen, numérique, texte intégral)
- Normalisation des scores pour la fusion hybride (Reciprocal Rank Fusion)
- Atelier pratique : Optimisation du taux de rappel (Recall) sur un corpus métier réel.

[Jour 2 - Après-midi]

Reranking et Post-processing

- Intégration de modèles de Re-ranking (Cross-Encoders comme BGE ou ColBERTv2)
- Gestion de la fenêtre de contexte et compression sémantique des prompts
- Réécriture adaptative de requêtes (Query Rewriting) pour corriger les ambiguïtés
- Détection et élimination des doublons sémantiques dans les résultats

Modélisation de Knowledge Graph

- Limites de la similarité sémantique : pourquoi le vecteur ne suffit plus pour les liens globaux
- Modélisation des entités, relations et événements stratégiques avec Neo4j
- Installation et sécurisation de la base de graphe en environnement local isolé
- Conception de schémas de graphes extensibles et compatibles avec le raisonnement IA

[Jour 3 - Matin]

Extraction et Construction du Graphe

- Extraction automatisée d'entités et de relations complexes par LLM locaux
- Maîtrise du langage Cypher pour l'interrogation et la manipulation de la connaissance
- Dédoublonnage et résolution d'entités pour la cohérence du graphe
- Validation de la structure du graphe par rapport au schéma métier défini
- Atelier pratique : Peuplement automatique d'un Knowledge Graph depuis le flux documentaire.

[Jour 3 - Après-midi]

Raisonnement Hybride GraphRAG

- Implémentation du GraphRAG : fusionner la recherche vectorielle et structurelle
- Navigation dans le graphe pour enrichir le contexte
- Gestion des schémas JSON et Function Calling pour des réponses structurées
- Algorithmes de recherche de chemins et détection de communautés pour le résumé global
- Atelier pratique : Développement d'un moteur de réponse liant des faits entre documents distants.

Orchestration avec LlamaIndex et LangChain

- Design de patterns d'orchestration : Router Query et Sub-question Query Engine
- Gestion de l'état et mémoire longue durée
- Observabilité des chaînes : traçabilité des étapes et gestion fine des logs d'exécution
- Découpage des requêtes complexes en sous-tâches parallélisables

[Jour 4 - Matin]

Agentic RAG et Autonomie

- Conception d'agents capables de planification et d'appels d'outils externes
- Boucles de réflexion pour minimiser les erreurs
- Génération automatisée de rapports multi-documents structurés selon des templates
- Gestion des conflits d'informations entre sources contradictoires
- Atelier pratique : Création d'un agent autonome.

[Jour 4 - Après-midi]

Évaluation Scientifique

- Framework RAGAS : mesures de Faithfulness, Answer Relevance et Context Precision
- Audit des performances : isoler les erreurs de recherche des erreurs de synthèse
- Création de bancs d'essais reproductibles avec des "Gold Datasets"
- Calcul du coût et de la latence par requête pour l'optimisation économique

IA Interprétable et Confiance

- Mécanismes de citations claires et vérification des sources en temps réel
- Affichage des scores de confiance et des chemins de pensée (Chain-of-Thought)
- Interface de correction humaine pour l'amélioration continue (Human-in-the-loop)
- Explicabilité des résultats : analyse de l'importance des documents récupérés

[Jour 5 - Matin]

Performance et Cache Sémantique

- Mise en œuvre du Semantic Caching pour réduire les temps de réponse et l'usage GPU
- Optimisation de la latence "Time To First Token" et du débit global
- Stratégies de pré-calcul et parallélisation des requêtes
- Gestion des files d'attente pour absorber les pics de charge

[Jour 5 - Après-midi]

LLMOps et Cycle de vie

- Monitoring des ressources hardware et détection de dérive sémantique
- CI/CD pour l'IA : tests de non-régression sémantique automatisés
- Gestion des versions de modèles et ré-indexation sans downtime
- Automatisation des boucles de feedback et collecte de métriques d'usage métier

Sécurité et Playbook de Production

- Protection contre les Prompt Injections et mise en place de Guardrails de sortie
- Contrôle d'accès au niveau granulaire et chiffrement On-Premise
- Élaboration du Playbook de production : procédures de backup et restauration
- Conformité RGPD en environnement souverain et isolation physique des données
- Atelier pratique : Audit complet, durcissement du système et validation du protocole de sécurité.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des

séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.