

Construire un RAG sur GCP

Support de cours — Vertex AI RAG Engine, Gemini, Cloud Run

Version 1.0 — octobre 2026

1. Pourquoi le RAG ?

Un LLM ne connaît que ses données d'entraînement. Le Retrieval-Augmented Generation (RAG) injecte au moment de la requête des passages pertinents issus de vos documents, ce qui réduit les hallucinations et permet de citer les sources.

Sur Google Cloud, Vertex AI RAG Engine gère l'ingestion (Cloud Storage, Google Drive), le découpage en chunks, le calcul des embeddings et l'indexation vectorielle. Le modèle Gemini interroge le corpus via un outil de retrieval.

2. Étapes clés

- Créer un corpus RAG et choisir le modèle d'embedding.
- Importer les fichiers (chunk_size, chunk_overlap).
- Interroger avec retrieval_query ou via un outil Gemini.
- Évaluer : précision du retrieval, fidélité de la réponse.

3. Comparatif des options de stockage vectoriel

Option	Gestion	Échelle	Cas d'usage
RagManagedDb	Entièrement managé	Moyenne	Démarrage rapide
Vertex AI Vector Search	Managé	Très grande	Production à fort volume
AlloyDB pgvector	Semi-managé	Grande	Données transactionnelles + vecteurs
Pinecone / Weaviate	Tiers	Grande	Multi-cloud

4. Bonnes pratiques

Chunks de 512 à 1024 tokens avec 10 à 20 % de recouvrement, filtrage par métadonnées, re-ranking des résultats et seuil de distance vectorielle pour écarter le bruit.