



Veille Technologique Quotidienne

Technique

18 septembre 2026 – Fréquence : Quotidienne (du Lundi au Vendredi)

Synthèse & Sujet du Jour

Milvus Distribué & OpenShift Data Foundation : Architecture Stockage Haute Performance pour le RAG Entreprise

Comment adosser une base vectorielle Milvus 2.4+ distribuée à Rook-Ceph et ODF sous OpenShift pour des requêtes sémantiques sous les 5 ms sur 10M+ de vecteurs.

1. Périmètre

Famille de veille : 2. Métiers, Compétences & Technologies.

Fréquence d'exécution : Quotidienne (du Lundi au Vendredi à 08h30).

Diffusion : [Radar Web Aperta](#) — [LinkedIn Aperta Scientia](#) — [X @ApertaScientia](#).

2. Sujet du Jour — Article Publié sur le Radar Web

Milvus Distribué & OpenShift Data Foundation : Architecture Stockage Haute Performance pour le RAG Entreprise

Dans les architectures RAG (Retrieval-Augmented Generation) d'entreprise, la scalabilité du pipeline s'arrête souvent là où le stockage vectoriel commence. Faire passer une base de connaissances métier de quelques milliers de documents à plus de 10 millions de vecteurs d'embeddings met sous tension critique la couche de persistance sous-jacente.

L'association d'une base vectorielle distribuée comme **Milvus 2.4+** et d'**OpenShift Data Foundation (ODF / Rook-Ceph)** offre une architecture découplée, souveraine et hautement disponible, capable de garantir des temps de réponse sémantique sub-5ms en environnement bare-metal.

1. Les limites des bases vectorielles monolithiques en production

Dans un déploiement RAG non découplé (base vectorielle standalone sur volume local) :

- **Goulot d'étranglement compute / storage :** Les nœuds de requêtes (*QueryNodes*) et les nœuds d'ingestion (*DataNodes*) se disputent les mêmes I/O disques et la même mémoire vive.
- **Absence de scalabilité élastique :** Impossible d'augmenter le débit de lecture sémantique indépendamment de la capacité de stockage brute des index HNSW/IVF-PQ.
- **Vulnérabilité aux pannes nœud :** La corruption ou la perte d'un nœud worker entraîne l'indisponibilité totale de l'index vectoriel et l'interruption des ServingRuntimes LLM associés.



2. L'architecture distribuée Milvus adossée à OpenShift Data Foundation

Le déploiement de Milvus sous OpenShift repose sur une séparation stricte des plans de données et de stockage :

- **Couche de persistance Ceph RBD / CephFS** : ODF fournit des volumes blocs haute performance (Ceph RBD) avec CSI provisionner dynamique pour stocker les index vectoriels et les journaux de segments persistants.
- **Découplage des flux avec Pulsar/Kafka et MinIO/S3** : L'ingestion des nouveaux embeddings transite par un broker d'événements distribué, tandis que le stockage objet ODF (Ceph RGW ou MinIO) conserve les snapshots immuables des collections vectorielles.
- **Isolation des QueryNodes bare-metal** : Les nœuds de calcul dédiés à la recherche vectorielle chargent les index en RAM sans bloquer les pipelines d'ingestion asynchrones, assurant une latence de similarité vectorielle inférieure à 5 ms.

3. Alignement Pédagogique & Compétences dans le Cours AS300

Chez Aperta Scientia, la conception d'usines IA d'entreprise intègre la maîtrise complète de la chaîne de données :

- **Cours AS300 (AI Platform Engineer — 399h)** : Déploiement de bases vectorielles distribuées (Milvus, Qdrant), interconnexion avec ODF et serving LLM KServe/vLLM sur clusters OpenShift bare-metal.
- **Cours AS200 (DevOps Platform Engineer)** : Administration et résilience du stockage logiciel avec OpenShift Data Foundation et monitoring granulaire sous Prometheus/Grafana.

4. Sources & Références Techniques

- **Milvus Distributed Architecture Documentation** : [__HREF_0__](#)
- **OpenShift Data Foundation (ODF) Storage Architecture** : [__HREF_0__](#)
- **Rook-Ceph Storage Operator for Kubernetes** : [__HREF_0__](#)

3. Impact sur les Cours Aperta Scientia

Cours AS200 — DevOps Platform Engineer

Cette veille alimente directement les modules d'ingénierie de plateforme et de sécurisation avancée de la chaîne de livraison logicielle :

- **Construction & CI/CD (D0288)** : Conception de pipelines déclaratifs Tekton, gestion des builds conteneurisés en environnement sans privilèges (*rootless*) et automatisation via Pipelines-as-Code.
- **Exploitation & Durcissement (D0380)** : Gestion multi-cluster, observabilité de la chaîne de livraison et politiques de conformité à l'échelle.
- **Sécurisation Conteneurs & Plateforme (D0430)** : Implémentation du contrôle d'admission (Sigstore Policy Controller, RHACS), signature keyless d'artefacts (RHTAS), génération de SBOM et analyse continue d'exploitabilité (TPA / VEX).



- **Certifications préparées :** Préparation opérationnelle aux examens **EX188** (Containers), **EX280** (OpenShift Administration) et **EX288** (OpenShift Development).

Cursus AS300 — AI Platform Engineer

La démarche de sécurisation de la supply chain logicielle est transposée à l'infrastructure d'inférence et aux pipelines MLOps :

- **Traçabilité des dépendances GPU & ML :** Audit de provenance et validation SLSA des images de base contenant les runtimes CUDA, ROCm et PyTorch.
- **Signature & Contrôle des Modèles :** Signature cryptographique des poids de modèles et des runtimes de serving vLLM / KServe avant déploiement en cluster.
- **Gouvernance & SBOM IA :** Inventaire automatisé des licences et détection proactive des vulnérabilités sur la pile MLOps (KubeFlow Pipelines, Ray, Triton).

4. Sources

- Red Hat Developer Blog — [Lien](#)
- OpenShift Documentation — [Lien](#)
- Kubernetes Community Documentation — [Lien](#)
- CNCF Cloud Native Landscape — [Lien](#)
- vLLM Project Documentation — [Lien](#)
- Sigstore Documentation — [Lien](#)
- SLSA Framework — [Lien](#)

Contacts & Communauté :

- Direction des Formations : learning@apertascientia.io
- Référent Handicap & Bien-être (Care) : care@apertascientia.io
- X (Twitter) : [@ApertaScientia](https://twitter.com/ApertaScientia)
- Communauté Discord : discord.gg/eDJzR5Vk85