**5. Partie IV – Déploiement et MLOps**

La dernière partie se concentre sur la mise en production et l’exploitation en conditions réelles.

* **Infrastructure et accélération** : comparaison des plateformes matérielles (GPU, TPU, CPU) et méthodes de parallélisation pour réduire le temps d’entraînement tout en conservant la précision.
* **Optimisation des modèles** : techniques de quantification et d’élagage pour réduire la taille des modèles et accélérer l’inférence, sans trop dégrader la performance.
* **Containers et services** : bonnes pratiques de conteneurisation avec Docker et Kubernetes, exposition des modèles via API REST, et intégration continue/déploiement continu (CI/CD) pour assurer un cycle de développement fluide.
* **Surveillance et maintenance** : suivi des performances en temps réel, détection de dérive des données, stratégies de réentraînement et de validation post-déploiement pour garantir la robustesse et l’éthique des applications.