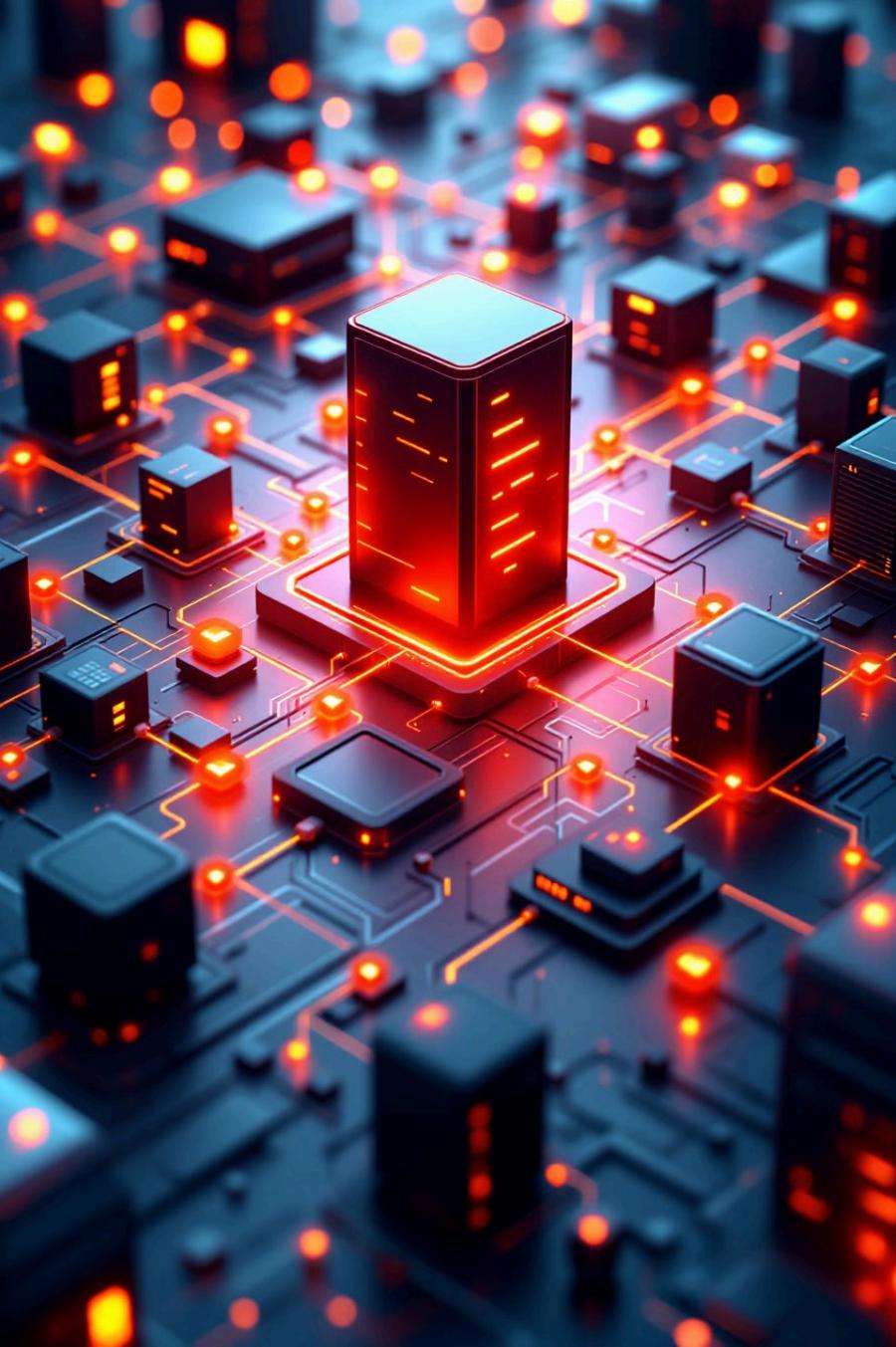


# Gestion et Déploiement de l'IA en Entreprise

Bienvenue à cette présentation sur la gestion et le déploiement de l'Intelligence Artificielle en entreprise. Nous explorerons en profondeur le cycle de vie d'un projet IA, du fine-tuning à l'hébergement, en passant par la mise à l'échelle, la maintenance continue et la conformité interne.

Cette présentation s'adresse aux ingénieurs, data scientists, responsables techniques et architectes IT, avec pour objectif de vous fournir une compréhension approfondie des défis et des meilleures pratiques dans ce domaine en constante évolution.





# Introduction : L'IA dans l'Entreprise Moderne

## ■ Intégration de l'IA

L'IA s'intègre dans l'entreprise moderne pour améliorer l'efficacité, l'innovation et la prise de décision.

## ■ Défis Majeurs

Complexité des modèles, coûts élevés, sécurité renforcée, besoin d'évolutivité et de contrôle.

## ■ Axes Abordés

Fine-tuning, optimisation, déploiement on-premise, mise à l'échelle, maintenance, conformité.

# Cycle de Vie des Projets IA en Entreprise

1

## Collecte et Préparation des Données

Rassemblement et nettoyage des données pertinentes pour l'entraînement du modèle.

2

## Entraînement et Validation

Développement du modèle et vérification de ses performances.

3

## Déploiement et Monitoring

Mise en production et surveillance continue des performances.

4

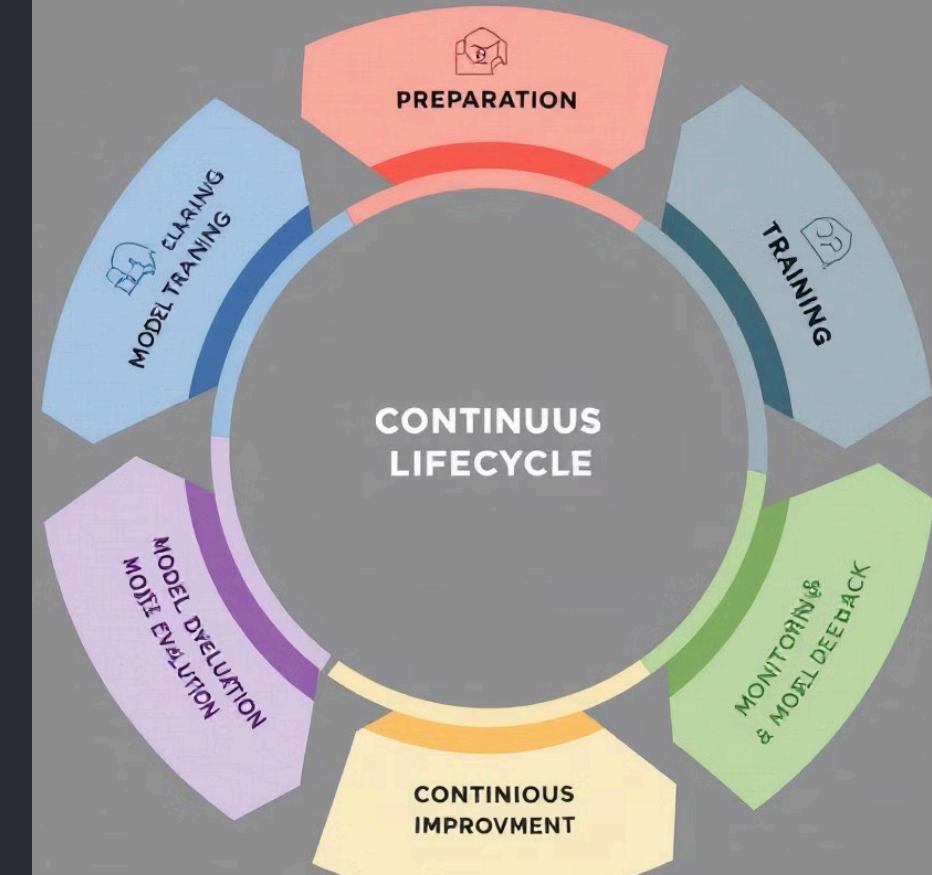
## Amélioration Continue

Optimisation et mise à jour régulière du modèle.

## AI PROJECT Lifecycle

### AI PROJECT LIFECYCLE

Data collection in model training, a model training in forr model evaluating, mcel evaluation, model, monitoring, and continuous improvises.



# Fine-tuning : Définition et Importance

## Qu'est-ce que le Fine-tuning ?

Le fine-tuning consiste à adapter un modèle pré-entraîné à un contexte spécifique, comme un domaine métier particulier ou des données internes à l'entreprise. Cette étape est cruciale pour optimiser les performances du modèle sur des tâches spécifiques.

## Pourquoi est-ce important ?

Le fine-tuning permet d'améliorer significativement la précision et la pertinence des prédictions du modèle dans le contexte de l'entreprise. Il réduit le temps et les ressources nécessaires par rapport à l'entraînement d'un modèle à partir de zéro.

## AI Model Optimization

AI Best of Model Optimizations	
 <b>Modeling</b> AI model optimization and upscaling to prepare for deployment and scalability and efficiency.	 <b>Model Rules</b> AI models' model techniques are optimized for use in production.
 <b>Model Optimization</b> How to load into the application, parameters and environment variables.	 <b>Change Research Optimization</b> Building a maintainable and clean codebase.
 <b>Sequences and iteration</b> Data consistency and memory management application.	 <b>Minimizing overhead</b> Efficiently and effectively managing resources in execution.
 <b>UX/FUI Design</b> How to minimize user interaction, files to consider before?	 <b>Cleaning noble collection</b> What is an encoder.
 <b>Baked Models</b> Autosetting options for targeting with day, the best and standardization.	 <b>Folded links</b> Reusable, and novel protection optimization.
 <b>UI/UX</b> Lastly there is a wills and wills and managing test for actions.	 <b>Direct live application operation</b> and reduce unnecessary storage.
 <b>Code Diction</b> The collection, protection and optimization.	 <b>UX proportion</b> Access here places the hypotheses and models and operation.
 <b>How to Model</b> Engage for optimizations performances compared using teachers and applications.	 <b>Code review optimization</b> Destitutes to reveal and to reflect in mode.

# Techniques de Fine-tuning et d'Optimisation



## Ajustement des Hyperparamètres

Optimisation des paramètres de configuration du modèle.



## Transfer Learning

Utilisation des connaissances acquises sur une tâche pour en améliorer une autre.



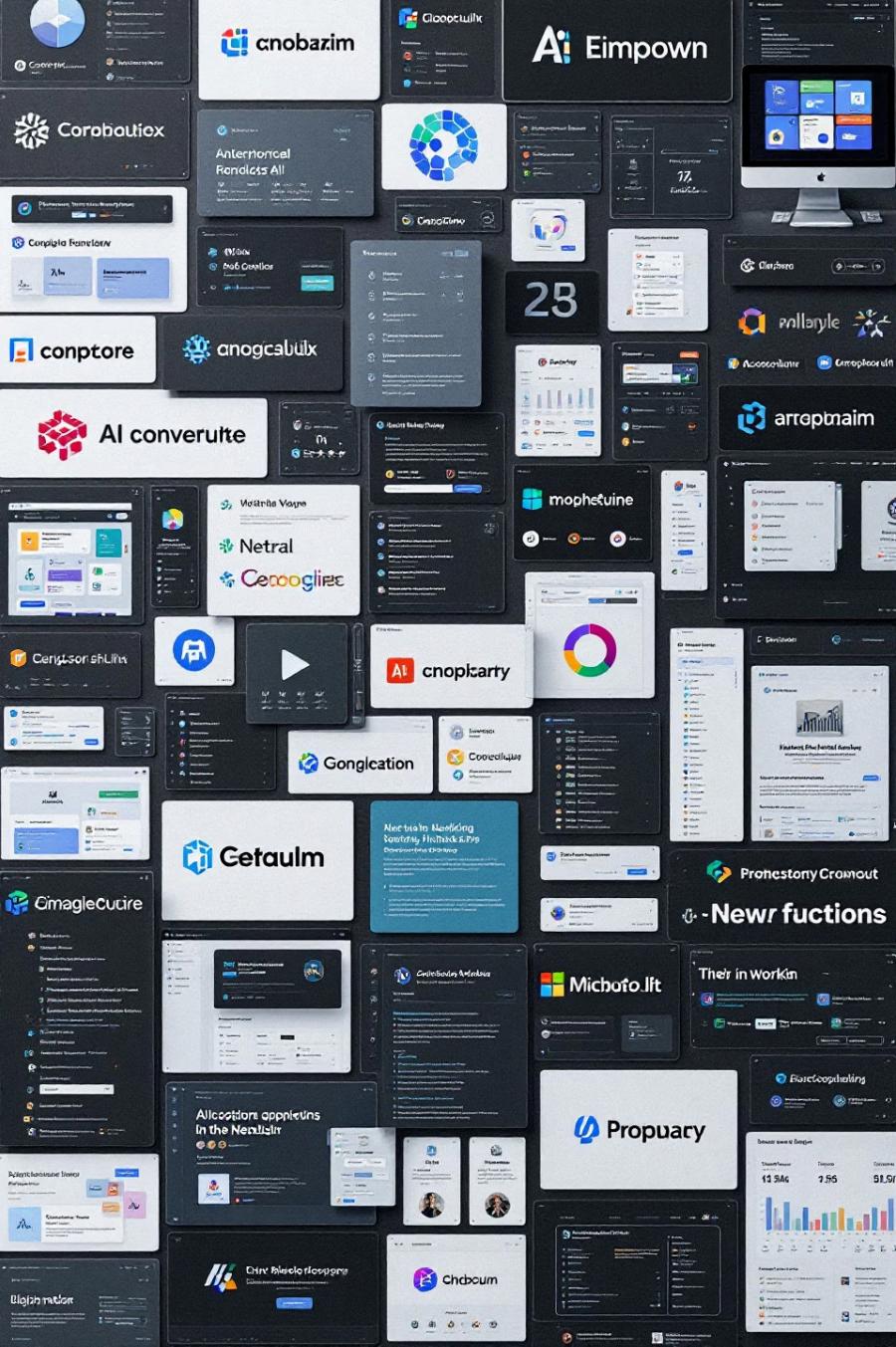
## Distillation de Modèles

Création de modèles plus légers à partir de modèles plus grands.



## Pruning et Quantization

Réduction de la taille et de la complexité du modèle.



# Outils et Frameworks pour le Fine-tuning

## Hugging Face Transformers

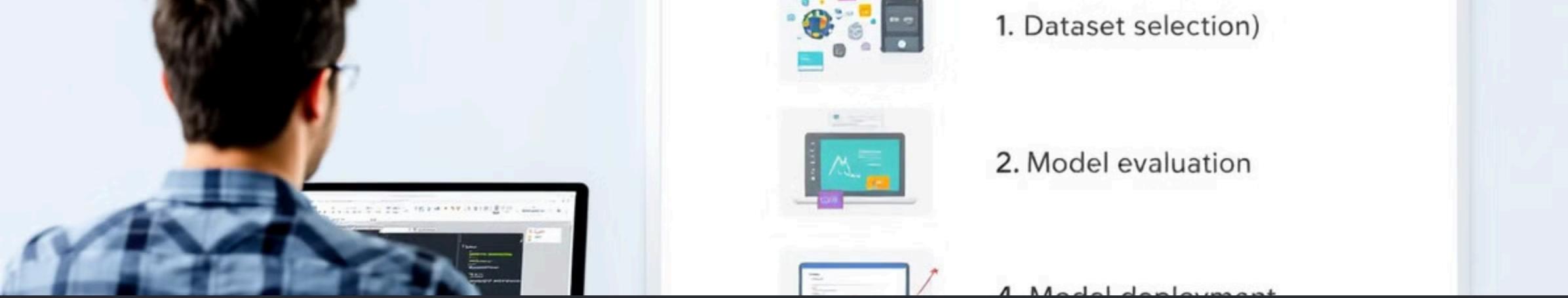
Bibliothèque open-source pour le traitement du langage naturel, offrant de nombreux modèles pré-entraînés et des outils de fine-tuning.

## TensorFlow et PyTorch

Frameworks de deep learning populaires avec des fonctionnalités avancées pour l'optimisation et le fine-tuning de modèles.

## Azure ML et Vertex AI

Plateformes cloud offrant des services de machine learning, y compris des outils pour le fine-tuning et le déploiement de modèles.



# Méthodologie pour le Fine-tuning



## Sélection du Dataset

1

Choisir des données représentatives du domaine d'application spécifique.

## Définition des Métriques

2

Identifier les indicateurs de performance pertinents pour évaluer le modèle.

## Validation Croisée

3

Utiliser des techniques de validation pour assurer la robustesse du modèle.

## Mise en Production

4

Déployer la version ajustée du modèle dans l'environnement de production.

# Étude de Cas : Fine-tuning pour la Classification de Documents

## Contexte

Une entreprise souhaite automatiser la classification de ses documents internes à l'aide d'un modèle de langage.

## Approche

Fine-tuning d'un modèle pré-entraîné sur un corpus de documents internes étiquetés.

## Résultats

Amélioration significative de la précision de classification par rapport au modèle générique.



# Déploiement On-premise : Pourquoi Choisir cette Option ?



## Contrôle Total des Données

Garantit que les données sensibles restent dans l'infrastructure de l'entreprise.



## Conformité Stricte

Facilite le respect des réglementations spécifiques à l'industrie ou à la région.



## Latence Réduite

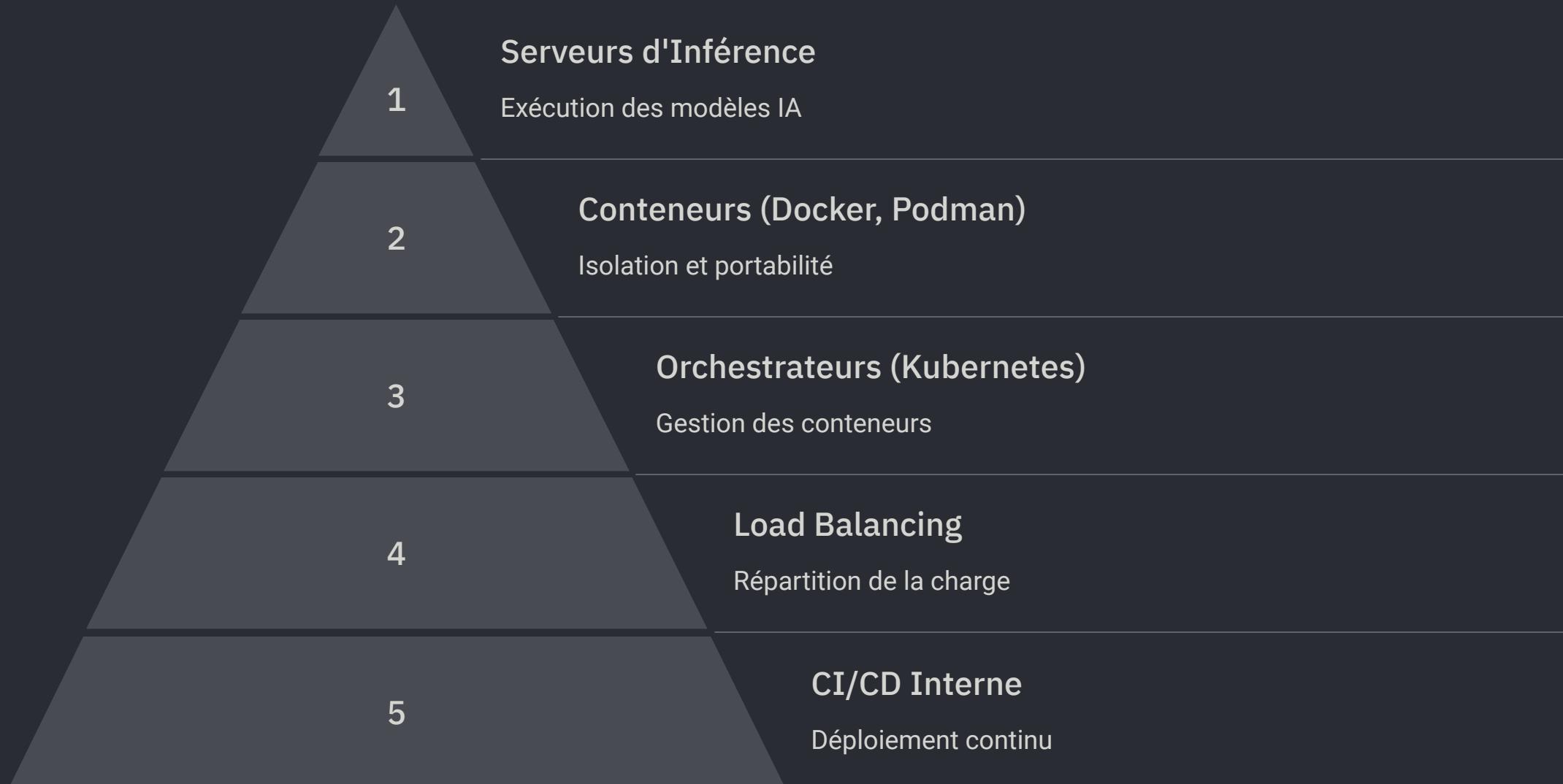
Minimise le temps de réponse pour les applications critiques.



## Intégration Interne

S'intègre plus facilement avec les systèmes et processus existants de l'entreprise.

# Architecture Type pour l'Hébergement On-premise





# Outils et Plateformes pour le Déploiement On-premise



## Kubernetes On-Prem

Orchestration de conteneurs pour le déploiement et la gestion de modèles IA.



## OpenShift

Plateforme d'entreprise basée sur Kubernetes pour le déploiement d'applications conteneurisées.



## MLflow

Plateforme open-source pour la gestion du cycle de vie complet des modèles ML.



## Kubeflow

Toolkit ML pour Kubernetes, facilitant le déploiement de workflows ML.



# Sécurité et Réseau pour le Déploiement On-premise

## Pare-feux

Protection contre les accès non autorisés et les menaces externes.

## Authentification et Autorisation

Contrôle d'accès granulaire aux ressources et aux modèles IA.

## Chiffrage des Données

Protection des données en transit et au repos pour garantir la confidentialité.

# Stratégies de Mise à Jour et de Rollback

1

## Planification

Préparation détaillée des mises à jour, incluant les tests et les scénarios de rollback.

2

## Déploiement Progressif

Mise à jour par phases pour minimiser les risques et détecter les problèmes tôt.

3

## Monitoring

Surveillance étroite des performances post-déploiement.

4

## Rollback Automatisé

Capacité de revenir rapidement à la version précédente en cas de problème.

model update:  
update and rollback



Rollback  
previous version



# Cas de Figure : Déploiement d'un Modèle NLP d'Assistance Interne

## Objectif

Déployer un modèle NLP pour gérer les FAQ et le support interne au sein du data center de l'entreprise.

## Défis

Intégration avec les systèmes existants, gestion de la confidentialité des données, et optimisation des performances.

## Solution

Utilisation de conteneurs Kubernetes pour le déploiement, avec une architecture microservices pour faciliter les mises à jour et le scaling.

# Mise à l'Échelle : Scaling Horizontal et Vertical

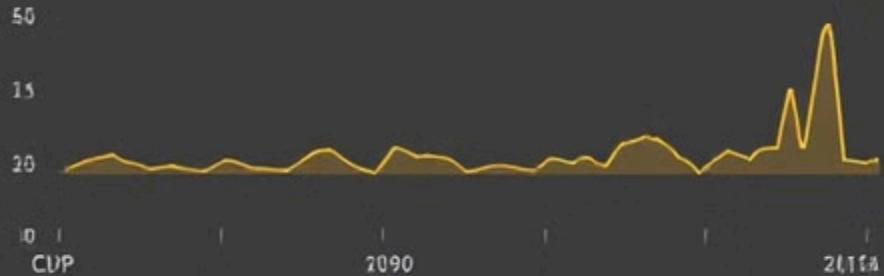
## Scaling Horizontal

Ajout de nouvelles instances ou serveurs pour répartir la charge.  
Idéal pour les applications distribuées et les microservices.

## Scaling Vertical

Augmentation des ressources (CPU, GPU, mémoire) d'un serveur existant. Adapté aux applications monolithiques ou nécessitant des ressources importantes.

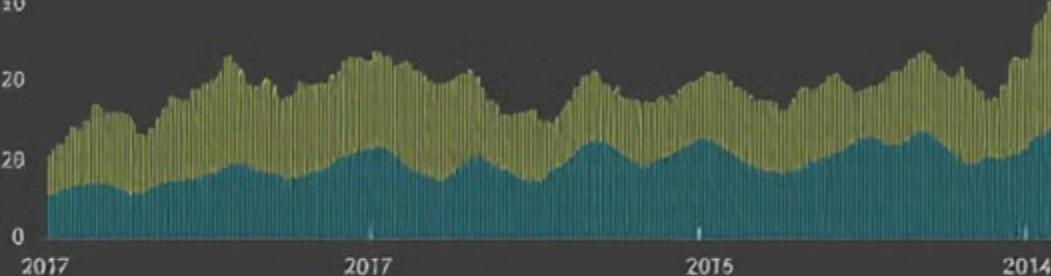
## Autoscaling (v)



CPU usage exceeded

Critical threshold exceeded

## Latency



Memory latency

Warning level exceeded

# Autoscaling et Monitoring



## Métriques Clés

Latence, taux d'erreur, consommation CPU/GPU, utilisation de la mémoire.



## Seuils d'Autoscaling

Définition de règles pour déclencher automatiquement l'ajout ou la suppression de ressources.



## Outils de Monitoring

Utilisation de plateformes comme Prometheus et Grafana pour la visualisation en temps réel.



## Alertes

Configuration d'alertes pour prévenir les équipes en cas de problèmes de performance.

# Thereiny AI Models Continuous Maintenance

# Maintenance Continue des Modèles IA

1

## Audits de Performance

Évaluation régulière des performances du modèle par rapport aux métriques définies.

2

## Détection de Dérive

Surveillance du data drift et du concept drift pour identifier les changements dans les données ou le contexte.

3

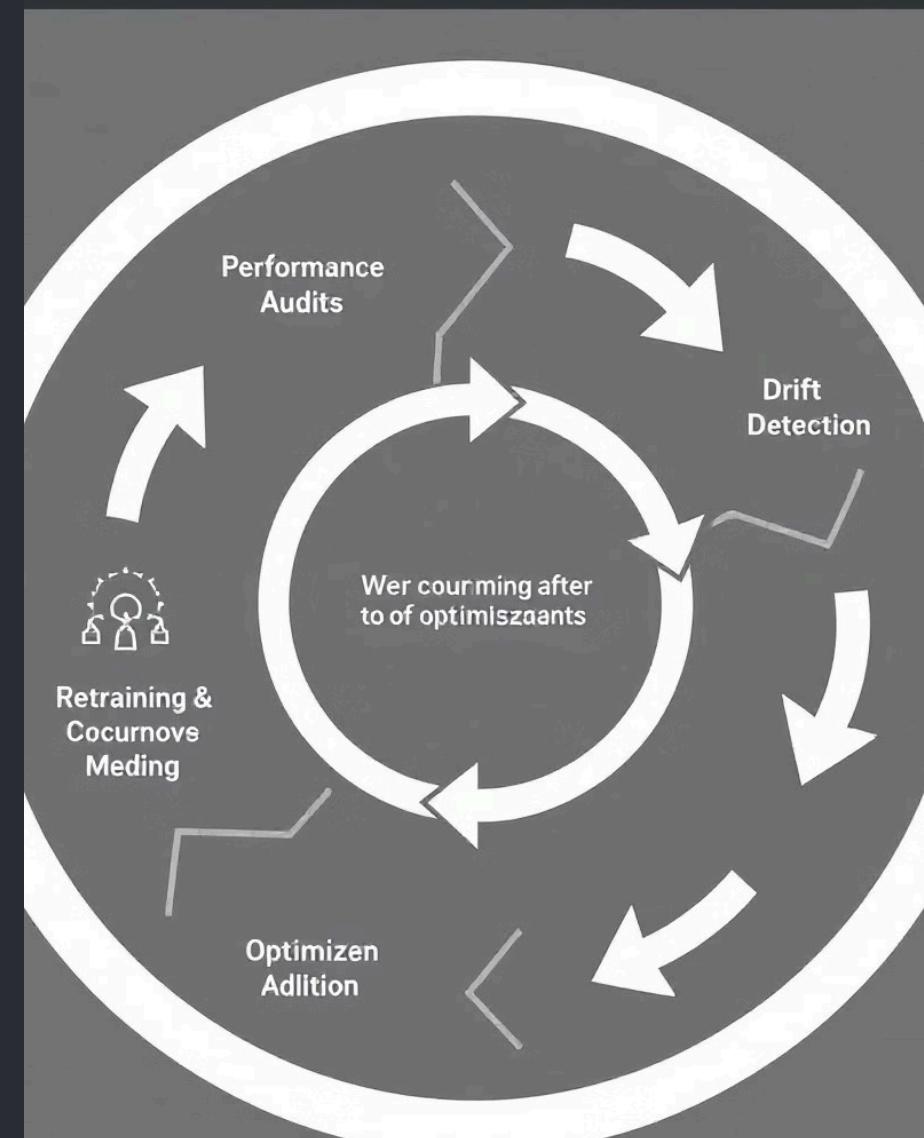
## Ré-entraînement

Mise à jour périodique du modèle avec de nouvelles données pour maintenir sa pertinence.

4

## Optimisation Continue

Ajustement des hyperparamètres et de l'architecture pour améliorer les performances.



This inhaue or continuous inainueance proress of tluifinution,  
reclaiting an tradel, angings meats of motel, contour and once  
tomding and in the maintenance.

# Logging et Traçabilité

## Journalisation Détailée

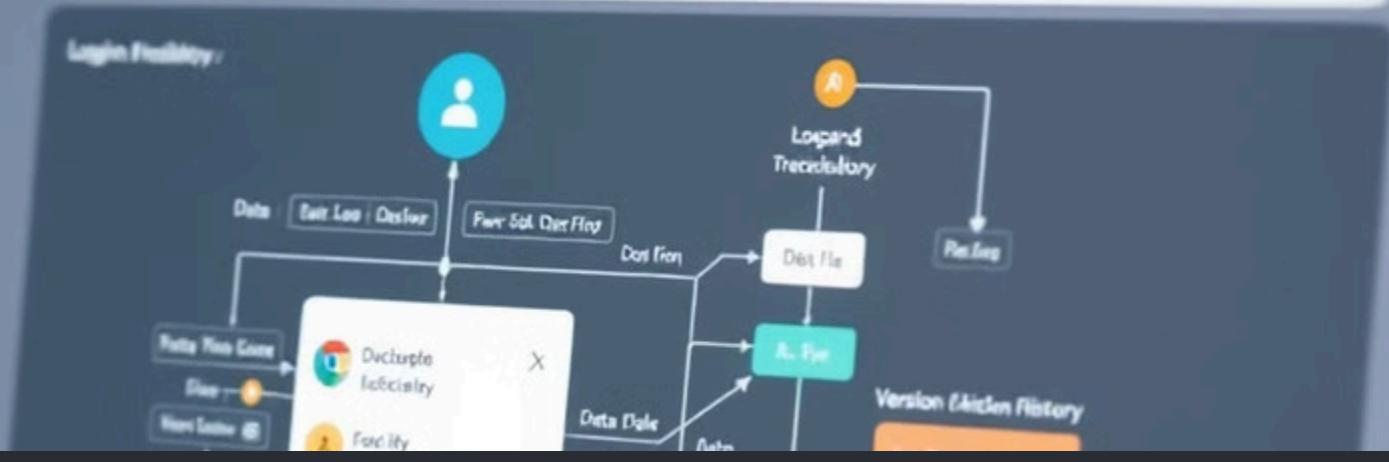
Enregistrement des entrées, sorties, et décisions du modèle pour chaque inférence.

## Historique des Versions

Suivi des changements apportés au modèle au fil du temps.

## Audit Trail

Documentation des accès et modifications pour la conformité et la sécurité.



# A/B Testing de Modèles IA

## Objectif

Comparer les performances de différentes versions d'un modèle en conditions réelles.

## Méthodologie

Déploiement parallèle de versions différentes du modèle, avec répartition aléatoire du trafic.

## Analyse

Évaluation des métriques de performance pour déterminer la version la plus efficace.



# Alertes et Observabilité



## Dashboards

Utilisation de Grafana pour visualiser les métriques clés en temps réel.



## Système d'Alerte

Configuration de PagerDuty ou Slack pour les notifications d'incidents.



## Analyse des Tendances

Suivi des métriques sur le long terme pour identifier les problèmes potentiels.



## Rapports Automatisés

Génération de rapports périodiques sur la santé et les performances des modèles.

# Conformité et Normes Internes

## ■ Politiques de Sécurité

Respect des normes de sécurité de l'entreprise pour la protection des données et des systèmes.

## ■ Confidentialité

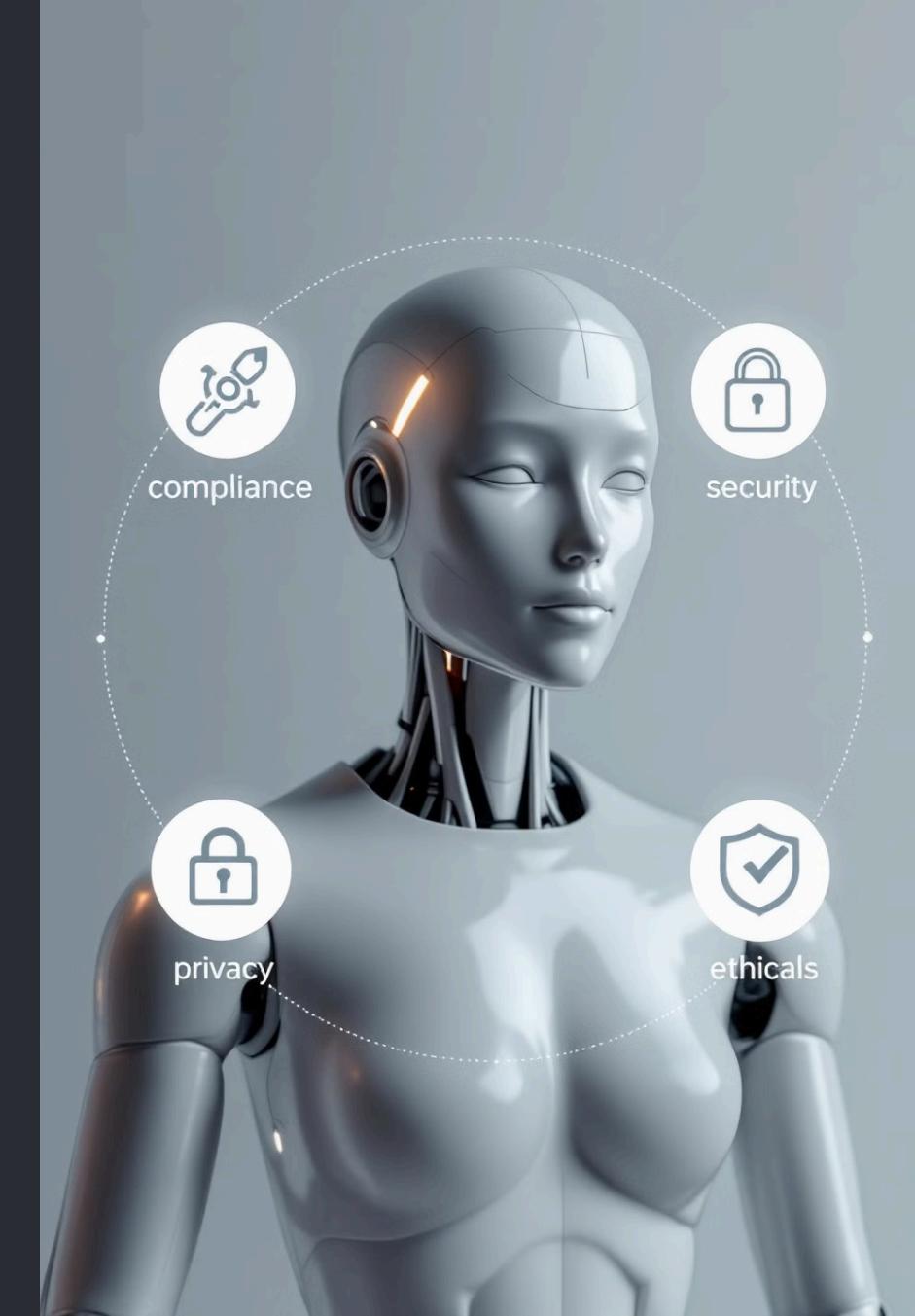
Mise en place de mesures pour garantir la confidentialité des données traitées par les modèles IA.

## ■ Conformité Réglementaire

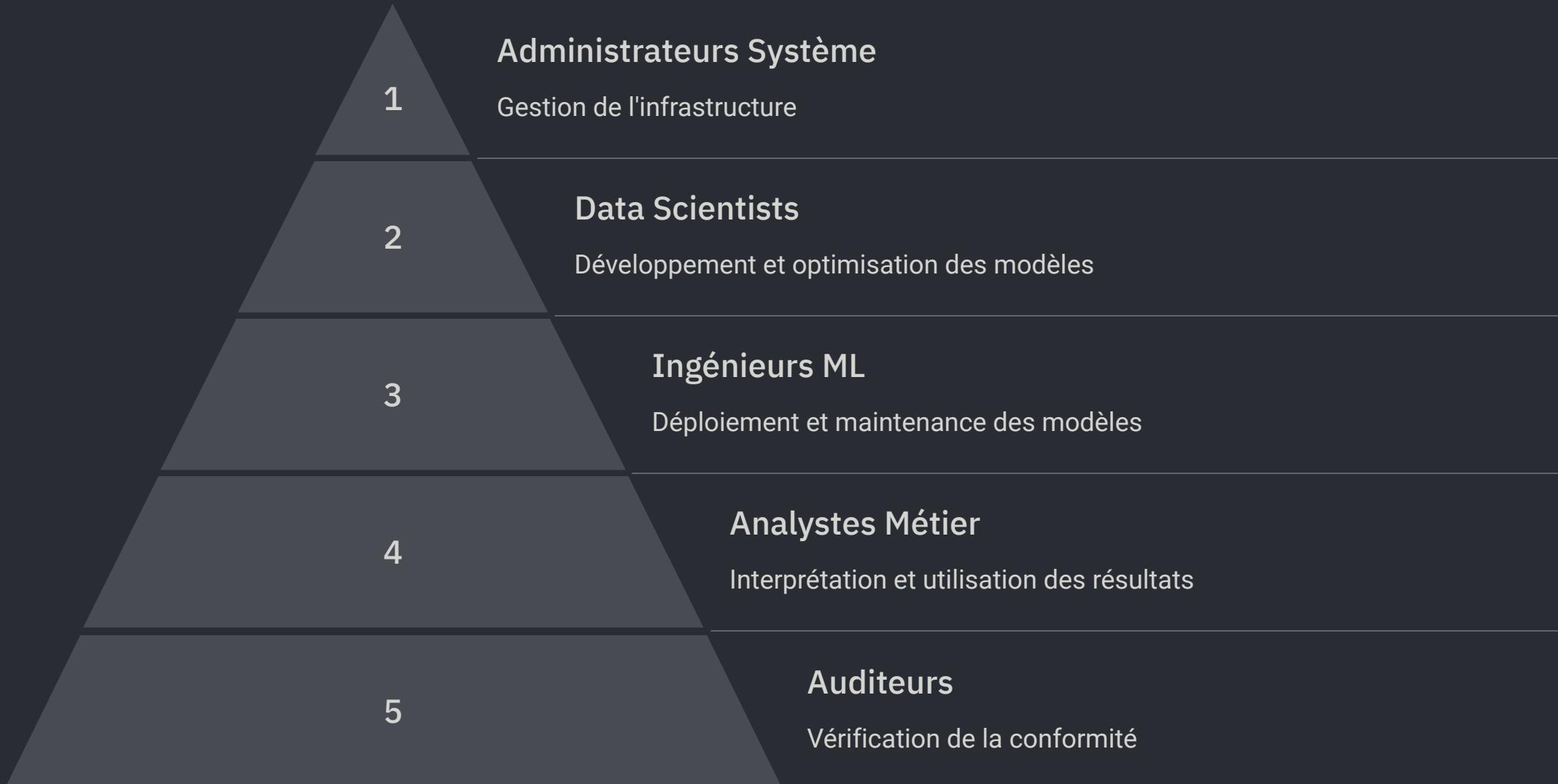
Adhésion aux réglementations comme le RGPD ou l'ISO 27001 pour la gestion des données.

## ■ Éthique de l'IA

Respect des principes éthiques dans le développement et l'utilisation de l'IA.



# Gestion des Accès et Responsabilités



# Séparation des Environnements

## Environnement de Développement

Espace dédié à l'expérimentation et au développement initial des modèles.

## Environnement de Test

Zone pour les tests approfondis et la validation des modèles avant production.

## Environnement de Production

Système isolé et sécurisé pour l'exécution des modèles en conditions réelles.

# SERVER ROOM



# Conservation des Artefacts et Reproductibilité

## Modèles

Stockage sécurisé des différentes versions des modèles IA.

## Datasets

Archivage des jeux de données utilisés pour l'entraînement et les tests.

## Logs

Conservation des journaux d'exécution et de performance.

## Configuration

Sauvegarde des paramètres et de l'environnement pour chaque version.

# Formation et Partage des Connaissances

1

## Programmes de Formation

Mise en place de formations régulières sur les technologies IA pour les équipes internes.

2

## Documentation

Création et mise à jour continue de guides techniques et de bonnes pratiques.

3

## Sessions de Partage

Organisation de présentations et de workshops pour partager les expériences et les apprentissages.

4

## Mentorat

Mise en place de programmes de mentorat pour faciliter le transfert de connaissances.



# Bonnes Pratiques : Standardisation du Cycle de Vie IA



## Définition des Standards

Établir des normes claires pour chaque étape du cycle de vie IA.

## Automatisation

Mettre en place des pipelines automatisés pour le développement et le déploiement.

## Contrôle de Version

Utiliser des outils de versioning pour les modèles, les données et le code.

## Revue de Code

Implémenter des processus de revue systématique pour assurer la qualité.



# Intégration de l'IA dans la CI/CD

## Tests Automatisés

Intégration de tests de performance et de qualité des modèles dans les pipelines CI/CD.

## Déploiement Continu

Mise en place de mécanismes pour déployer automatiquement les nouvelles versions des modèles après validation.

## Monitoring Intégré

Inclusion de métriques de performance des modèles dans les dashboards de monitoring existants.

# MLOps : Fusion du DevOps et de l'IA



## Automatisation

Automatisation des processus de bout en bout, de l'entraînement au déploiement.



## Collaboration

Facilitation de la collaboration entre data scientists, développeurs et opérations.



## Monitoring Continu

Surveillance en temps réel des performances et de la santé des modèles.



## Itération Rapide

Capacité à rapidement itérer et améliorer les modèles en production.





## Conclusion : Points Essentiels

### Cycle Complet

Du fine-tuning au déploiement on-premise, en passant par la mise à l'échelle et la maintenance.

### Approche Intégrée

Importance d'une approche holistique intégrant développement, opérations et conformité.

### Amélioration Continue

Nécessité d'un processus d'amélioration et d'optimisation constant des modèles et des pratiques.

### Collaboration

Importance de la collaboration entre équipes Data, Dev et Ops pour le succès des projets IA.