



Architectures de solutions IA

IA007

Durée: 3 jours

Public :

Administrateurs systèmes, ingénieurs DevOps, architectes infrastructure, administrateurs de données, ingénieurs systèmes déployant des solutions IA en production

Objectifs :

Maintenir et superviser des solutions IA en production, configurer des outils de surveillance et d'alertes, gérer les cycles de vie des modèles, administrer les infrastructures IA, assurer la sécurité et la conformité des déploiements

Connaissances préalables nécessaires :

Expérience en administration systèmes, notions de base en réseaux et bases de données.

Programme :

Modèles architecturaux pour l'IA

Comprendre les modèles fondamentaux des architectures IA et leurs cas d'usage.
Modèles batchs, temps-continu, presque temps réel : architectures Lambda, Kappa, Delta
Architecture événementielle pour l'IA : historisation des événements, séparation commandes/requêtes, CQRS, traitement en flux
Modèles microservices pour l'IA : découverte des services, disjoncteur, cloisonnement
Architecture sans serveurs pour l'IA : FaaS, pilotage par événements, élasticité automatique
Modèle edge computing : apprentissage de proximité en mode fog computing, apprentissage fédéré, inférence distribuée

Atelier : Analyse comparative des modèles - sélection d'architecture pour 3 cas d'usage différents (recommandation temps réel, analyse batchs, IoT de périphérie)

Architecture des chaînes de données et ML

Concevoir les architectures de traitement et d'ingestion des données pour l'IA.
Architecture de stockage : comparaison lac de données, entrepôts, lac pour l'IA
Chaînes ETL/ELT optimisées pour l'apprentissage automatique
Architecture de diffusion en continu : Apache Kafka, Pulsar, Kinesis - comparaison et choix
Magasins de caractéristiques (feature stores) : architecture centralisée vs décentralisée, en ligne vs hors ligne
Maillage de données (data mesh) et architectures distribuées pour l'IA
Versioning des données et reproductibilité : DVC, Delta Lake, Apache Iceberg

Atelier : Conception d'une chaîne de données ML - architecture lambda avec Kafka, magasin de caractéristiques, versioning, surveillance de qualité



Architecture d'infrastructure et dimensionnement

Dimensionner et concevoir l'infrastructure pour les charges de travail IA.
Dimensionnement GPU, CPU : critères de choix, coût/performance
Architecture multi-cloud et cloud hybride pour l'IA
Stratégies de mise à l'échelle : mise à l'échelle automatique, équilibrage de charge, mutualisation des ressources
Architecture haute disponibilité : redondance, basculement, reprise après sinistre
Optimisation des coûts : instances ponctuelles, capacité réservée, planification intelligente
Architecture edge-to-cloud : synchronisation, cohérence, optimisation de latence

Atelier : Dimensionnement d'une infrastructure IA - calcul des besoins, architecture haute disponibilité, stratégie multi-cloud, estimation des coûts

Architecture de déploiement et distribution de modèles

Concevoir les architectures de déploiement et de service des modèles ML.
Modèles de déploiement : bleu-vert, canari, tests A/B, mode fantôme
Architecture de distribution de modèles : synchrone vs asynchrone, par lots vs temps réel
Orchestration de modèles : méthodes d'ensemble, cascade, routage intelligent
Architecture multi-modèles : registre de modèles, gestion des versions, tests A/B
Optimisation des inférences : traitement par lots, mise en cache, quantification, élagage
Architecture d'inférence de périphérie : TensorFlow Lite, ONNX Runtime, NVIDIA Triton

Atelier : Conception d'une architecture de distribution - déploiement multi-modèles, équilibrage de charge, optimisation latence, surveillance des performances

Intégration dans les architectures SI existantes

Intégrer les solutions IA dans les systèmes d'information existants.
Modèles d'intégration : API-first, piloté par événements, files de messages
Architecture de passerelle d'APIs pour l'IA : limitation de débit, authentification, surveillance
Intégration avec systèmes hérités : adaptation de modèles d'architecture, couches anti-corruption
Architecture ESB et microservices : choix et compromis
Gestion des transactions et cohérence dans les systèmes IA
Impact sur les architectures existantes : migration de modèles d'architecture, coexistence

Atelier : Conception d'une stratégie d'intégration - passerelle API, intégration de modèles d'architecture, migration progressive, gestion des dépendances

Architecture de sécurité et gouvernance

Concevoir la sécurité et la gouvernance au niveau architectural. Modèles d'architecture sécurisée.
Architecture de confiance pour l'IA : identité, périphériques, réseau, données
Chiffrement et protection des données : au repos, pendant les transferts, pendant les calculs
Architecture de suivi et auditabilité : journalisation, traces, métriques
Modèles de gouvernance : politique de gouvernance en tant que code, conformité automatisée
Architecture multi-tenant : isolation, sécurité, performance

Atelier : Conception d'une architecture sécurisée - modélisation des menaces, architecture zéro confiance, automatisation de la conformité, surveillance de sécurité



Performance et optimisation architecturale

Optimiser les performances des architectures IA.

Métriques architecturales : latence, débit, disponibilité, évolutivité

Modèles d'optimisation : stratégies de mise en cache, mutualisation des connexions, traitement asynchrone

Architecture distribuée : fragmentation, partitionnement, distribution de charge

Optimisation réseau : CDN, mise en cache de périphérie, compression, optimisation de protocole

Surveillance et observabilité : APM, traçage distribué, ingénierie du chaos

Planification de capacité et tests de performance

Atelier : Optimisation d'une architecture existante - identification des goulots d'étranglement, stratégies d'optimisation, surveillance avancée, tests de charge

Architecture cloud-native et DevOps

Concevoir des architectures cloud-native pour l'IA.

Modèles natifs cloud : application 12-factor, conteneurisation, orchestration

Architecture Kubernetes pour l'IA : opérateurs, ressources personnalisées, planification GPU

CI/CD pour l'IA : GitOps, infrastructure en tant que code, tests automatisés

Architecture multi-régions : réplication de données, optimisation de latence, reprise après sinistre

Maillage de services et observabilité : Istio, Linkerd, suivi des traces distribuées

Architecture en tant que code : Terraform, Pulumi, modèles ARM

Atelier : Conception d'une architecture cloud-native complète - Kubernetes, CI/CD, maillage de services, multi-régions, infrastructure en tant que code

Études de cas et revue d'architecture

Appliquer les concepts à des cas réels et effectuer des revues d'architecture.

Architecture pour différents domaines : e-commerce, finance, santé, IoT

Revue d'architectures existantes : identification des anti-modèles, recommandations

Compromis architecturaux : performance vs coût, sécurité vs utilisabilité

Enregistrements de décisions architecturales (ADR) : documentation des choix architecturaux

Présentation et défense d'architectures

Méthodologies d'architecture : TOGAF, modèle C4, arc42

Atelier final : Conception complète d'une architecture IA pour un cas d'usage complexe - documentation ADR, présentation à un comité d'architecture, revue par les pairs