CERTIFICATION

DATA ARCHITECT

MLOPS, GENAI, DATA SCIENCE

FFCCERTTERXBD01	PRIX : 12 500 €	DURÉE : 26 JOURS	ÉLIGIBLE CPF

Pauses et déjeuners offerts

L'architecte de données (data architect) conçoit les systèmes de gestion de données massives et complexes. Il participe à la mise en production d'applications d'intelligence artificielle, contribue au passage à l'échelle et assure le déploiement des modèles. Sa mission MLOps (Machine Learning Operations) est de planifier les outils et les pratiques pour automatiser et optimiser les phases de développement, de test, de déploiement et de supervision des modèles d'apprentissage machine.

Cette activité MLOps devient plus cruciale que jamais avec le passage d'une phase R&D à la phase actuelle d'industrialisation massive et de génération de contenus avec l'IA générative (GenAI). À mesure que les entreprises intègrent des solutions d'intelligence artificielle dans leurs opérations quotidiennes, la gestion efficace du cycle de vie des modèles devient une priorité stratégique. L'IA générative permet de créer des contenus originaux et offre des opportunités innovantes mais nécessite un déploiement rigoureux pour garantir la performance, la précision et la fiabilité des modèles déployés.

Ces modèles peuvent évoluer rapidement. La créativité doit être équilibrée avec la précision des sciences des données (data science) et la conformité avec la réglementation . Le data architect assure non seulement la qualité et la cohérence des sorties générées, mais répond aussi rapidement aux nouvelles exigences, aux nouveaux algorithmes et au retour des utilisateurs. Il permet ainsi aux organisations de créer, de déployer et de gérer efficacement des modèles d'intelligence artificielle à grande échelle, tout en garantissant leur intégration dans les systèmes de données existantes, leur fiabilité, leur performance et leur sécurité.

VOUS ÊTES

Ingénieur ou chef de projet souhaitant développer des compétences dans le domaine de conception et déploiement de systèmes de données massives pour évoluer vers un poste tel que celui d'architecte des données (architecte data), d'ingénieur MLOps ou de responsable projet data.

Une connaissance de base en mathématiques appliquées et une bonne expérience de la conception technique d'applications ou de SGBD sont indispensables pour suivre avec profit cette formation (voir la formation MOOC Fondamentaux pour Big Data).

OBJECTIFS

- Concevoir une architecture de données et de traitements robuste et évolutive, y compris la modélisation, la génération des données et le stockage des données
- Intégrer les solutions d'IA dans des architectures de données existantes ou dans des architectures en conception (by Design) avec une approche frugale
- Produire des pipelines de données efficaces pour l'entraînement, le déploiement et la gestion des modèles de science de données et d'intelligence artificielle
- Gérer le cycle de vie des modèles, la supervision des performances et l'automatisation des processus de déploiement
- Limiter les risques dans la conception, le déploiement et l'exploitation de l'architecture de données

ÉVALUATION ET CERTIFICATION

Contrôle des acquis et des savoir-faire au travers de travaux individuels et de groupe.
Évaluation du mémoire professionnel basé sur un projet individuel soutenu devant un jury.

Les participants ayant suivi le parcours avec succès obtiennent la certification « RS6017- Data Science : analyse et gestion de grandes masses de données » (de niveau 7) de Télécom Paris.

PROGRAMME

Introduction

L'analyse exploratoire des données
- Explorer les données
- Traitement préalable des données : assurer la complétude, gérer les outliners, nettoyage, normalisation, atténuation, discrétisation
- Les données en tant qu'actifs de l'entreprise

Architecture d'une application IA
- Besoins, objectifs et contraintes
- Approche monolithique vs microservices
- Intégration de la sécurité dans le cycle de vie du développement applicatif

Enjeux de l'IA
- Risques juridiques et éthiques
- Risques produit des applications de science de données et IA
- Accompagnement du changement pour une transformation numérique avec l'IA
- Définir une stratégie de qualité de données (data quality)

Systèmes de traitement de données massives
- Applications cloud-natives
- Cadrer un projet de données intensives pour la science de données et l'IA
- Intégration de la sécurité dans le cycle de vie du développement logiciel Mapreduce
- Au-delà de mapreduce : spark, stratosphere
- Message passing interface
- Calculs distribués sur des graphes : graphlab, pregel, giraph
- Infrastructure et analyse continue
- Optimisation et déploiement continu de modèles

Apprentissage machine : de la théorie aux design patterns
- Produire un modèle d'apprentissage automatique
- Apprentissage supervisé, deeplearning et IA générative : éléments de la théorie de vapnik chervonenkis, support vector machines et arbres de décision
- Apprentissage en ligne
- Apprentissage statistique distribué
- Techniques d'échantillonnage
- Variables latentes et clustering
- Chaînes de markov cachés
- Réseaux de neurones, apprentissage par renforcement et traitement d'images
- Design patterns applicatifs et modèles

Traitement du langage naturel et Transformers
- Tokenisation
- Marquage de partie de discours
- Représentation de document
- Word embeddings
- Transformers, llms et infrastructures de calcul

Génération de données
- Retrieval augmented generation
- Fine tuning à partir de modèles pré-entraînés
- Génération des données visuelles
- Déployer un modèle d'IA générative

Stockage de données massives
- Sgbd relationnels distribués classiques
- Systèmes de fichiers distribués hdfs/gfs
- Stockage à grande échelle Stockage clés-valeurs par table de hachage distribuée (dynamo)
- Stockage par arbre distribué (bigtable et hbase)
- Systèmes newsql (google spanner, sgbd en mémoire, mysql cluster)

Apprentissage distribué : fouille de Graphes
- Distribution d'algorithmes d'indexation, d'apprentissage et de fouille
- Index inversé
- Factorisation de matrice
- Échantillonnage
- Pagerankk

Synthèse et conclusion

NOUVEAU PROGRAMME

ATELIER

FAISABLE À DISTANCE

EN

RÉALISABLE EN ANGLAIS

RESPONSABLE(S)

Louis JACHIET
Louis Jachiet est Maître de conférences en informatique à Télécom Paris au sein de l'équipe DIG (Data, Intelligence and Graphs). Il s'intéresse particulièrement à l'algorithmique, aux bases de données, aux langages de programmation et à la logique.

Yann BALGOBIN
Yann Balgobin est titulaire d'un doctorat de Télécom Paris. Son expertise porte sur l'économie numérique et ses enjeux : protection des données personnelles, économie des plateformes, enjeux socioéconomiques du numérique. Il est également responsable pédagogique des domaines « Management de la transformation numérique » et « Intelligence Artificielle et science des données » à Télécom Paris Executive Education.

CERTIFICATION DÉLIVRÉE PAR

TELECOM Paris

IP PARIS