Architecture des données : stockage et accès

Appréhender et prendre en main les nouvelles architectures de données : Hadoop, NoSQL, Spark

Stage pratique
Data Science

Code :
DSARC

Durée :
3 jour(s) (21,00 heures)

Exposés : **50 %**
Cas pratiques : **40 %**
Echanges d'expérience : **10 %**

Inter-entreprises :
Prochaines sessions disponibles sur notre site web.
Tarif : 2 500,00 € HT / participant

Intra-entreprise :
Tarifs et dates sur demande.

DESCRIPTION

Si les algorithmes de Machine Learning ont connu des avancées majeures ces dernières années, c'est avant tout grâce à la quantité d'information disponible pour les entraîner. Accumuler toute cette donnée, la traiter, et la rendre disponible sont les enjeux principaux du mouvement Big Data.

Au cours de cette formation, nos consultants mettent à disposition les connaissances issues de leurs retours d'expériences auprès de nos clients, et vous font découvrir les bases des architectures permettant de répondre à ces enjeux de stockage et d'accès.

OBJECTIFS PEDAGOGIQUES

- Découvrir les notions centrales de stockage de données
- Préciser les enjeux des nouvelles architectures de données (Hadoop, NoSQL, Spark), et positionner leurs usages au sein de l'univers Big Data
- Manipuler ces technologies et les bases de données de façon conjointe, pour mener à bien des analyses efficaces

PUBLIC CIBLE

- Analyste
- Statisticien
- Développeur

PRE-REQUIS

- Notions de programmation sur la base d'un langage quelconque
- Manipulation basique de la ligne de commande Linux

METHODE PEDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation.

PROFIL DES INTERVENANTS

Cette formation est dispensée par un·e ou plusieurs consultant·es

d'OCTO Technology ou de son réseau de partenaires, expert·es reconnus des sujets traités.

Le processus de sélection de nos formateurs et formatrices est exigeant et repose sur une évaluation rigoureuse leurs capacités techniques, de leur expérience professionnelle et de leurs compétences pédagogiques.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique.

Afin de valider les compétences acquises lors de la formation, un formulaire d'auto-positionnement est envoyé en amont et en aval de celle-ci.

En l'absence de réponse d'un ou plusieurs participants, un temps sera consacré en ouverture de session pour prendre connaissance du positionnement de chaque stagiaire sur les objectifs pédagogiques évalués.

Une évaluation à chaud est également effectuée en fin de session pour mesurer la satisfaction des stagiaires et un certificat de réalisation leur est adressé individuellement.

PROGRAMME PEDAGOGIQUE DETAILLE

Jour 1

MODULE 1 : LES FONDAMENTAUX ET LA STRATÉGIE DE DONNÉES

Avant d'aborder les outils et les technologies, nous nous concentrons sur les concepts essentiels, les principes de conception et la valeur métier.

- Introduction : rappel historique sur l'évolution de la donnée (Silos /Data Warehouse / Data Lake / Lakehouse)
- Le rôle de l'Architecte Data : différences avec le Data Engineer et le Data Scientist
- CAP Theorem & PACELC : comment les systèmes distribués équilibrent cohérence, disponibilité et latence
- OLTP vs OLAP : comprendre pourquoi il est essentiel de séparer les charges de travail transactionnelles (OLTP) et analytiques (OLAP)

MODULE 2 : MODÉLISATION DE LA DONNÉE (LE COEUR DU MÉTIER)

La modélisation des données a un coût : apprenez à structurer vos données pour éviter les erreurs coûteuses.

- Modélisation relationnelle : 3NF (Troisième forme normale)pour éviter les redondances et anomalies
- Modélisation dimensionnelle (Kimball) : schémas en étoile et en flocon, et gestion des changements lents (SCD)
- Data Vault 2.0 : concevoir des entrepôts de données d'entreprise à la fois agiles et auditables
- Modélisation NoSQL :
  - Document (MongoDB), Clé-Valeur (Redis), Colonne (Cassandra), Graph (Neo4j)
  - Patterns d'accès vs normalisation

Jour 2

MODULE 3 : PARADIGMES D'ARCHITECTURE MODERNE

Les clés pour assembler les briques et former un écosystème fonctionnel.

- Data Warehouse Cloud : Snowflake, BigQuery, Redshift. Architecture Compute/Storage découplé
- Data Lake : Stockage objet (S3, ADLS), formats de fichiers (Parquet, Avro, Delta Lake, Iceberg)
- Lakehouse : le meilleur des deux mondes (ACID directement sur le lac de données). Focus sur Databricks/Delta ou Apache Iceberg ou Ducklakehouse
- Architecture Lambda vs Kappa : gestion du temps réel et du batch
- Data Mesh & Data Fabric :
  - Décentralisation (Domain-Driven Design appliqué à la data)
  - La donnée comme produit (Data as a Product)
  - Gouvernance fédérée

MODULE 4 : INTÉGRATION ET PIPELINES DE DONNÉES

Assurer la circulation des données de la source vers la destination.

- ETL vs ELT : Pourquoi le "T" (Transformation) se déplace vers la fin ?
- Batch Processing : Orchestration avec Apache Airflow, Prefect ou Dagster
- Streaming & Event-Driven : Apache Kafka, Pulsar. Concepts de Producers/Consumers, Topics, partitions
- CDC (Change Data Capture) : capter les événements en temps réel depuis les bases de production

Jour 3

MODULE 5 : GOUVERNANCE, QUALITÉ ET SÉCURITÉ

Sans maîtrise, la puissance n'est rien.

- Data Quality : les 6 dimensions de la qualité. Outils de test (Great Expectations, dbt tests)
- Data Catalog & Lineage : Savoir d'où vient la donnée (DataHub, Alation, Collibra)
- Sécurité et Conformité : RBAC, ABAC, chiffrement, anonymisation, RGPD/GDPR
- FinOps Data : contrôler les coûts du Cloud

MODULE 6 : ARCHITECTURE POUR L'IA ET LE MACHINE LEARNING

Créer les conditions de succès pour les Data Scientists et l'IA générative.

- Feature Stores : centraliser les variables pour le ML (Feast, Hopsworks)
- Architecture MLOps : intégration du cycle de vie des modèles (Entraînement, Déploiement, Monitoring)
- Architecture pour l'IA Générative (LLM) :
  - Vector Databases : Pinecone, Weaviate, pgvector (pour le RAG - Retrieval Augmented Generation)
  - Gestion des données non structurées (PDF, Images, Audio)

- Pipelines d'ingestion pour LLM (LangChain/LlamaIndex)

Accessibilité

L'inclusion est sujet important pour OCTO Academy.
Nos référent·es sont à votre disposition pour faciliter l'adaptation de votre formation à vos besoins spécifiques.
Pour les contacter : academy.accessibilite@octo.com