Fiche du cours

55 h

Titre :

BGD300 - Big Data Engineering (Spark, Kafka, Lakehouse) – Python

Description :

Formation pratique à l'ingénierie Big Data en Python : ingestion de données (batch & streaming), traitements distribués avec PySpark, messagerie Kafka, stockage colonne Parquet/Delta, qualité des données (Great Expectations), tests (pytest), Docker et aperçu d'orchestration (Airflow/Dagster). Du fichier brut ou d'un flux temps réel jusqu'à un dataset propre, documenté et exploitable.

Objectifs :

- Comprendre l'architecture Big Data (batch vs streaming, parallélisme, formats colonnes).*
- Manipuler PySpark DataFrame/SQL et optimiser (partitions, joins, cache, lecture/écriture efficaces).*
- Mettre en place une ingestion batch (CSV/JSON → Parquet/Delta) et un pipeline streaming depuis Kafka*.
- Appliquer des contrôles de qualité (schémas, règles, rapports) et des tests orientés données*
- Conteneuriser la stack (Spark/Kafka) et livrer un README exécutable avec scripts de lancement.*

Chapitres :

1. Fondations Big Data & Setup : concepts clé, formats (CSV/JSON/Parquet/Delta), environnement Python propre, structure projet*
2. PySpark – bases : DataFrame, opérations courantes, Spark SQL, bonnes pratiques d'écriture*
3. Performance PySpark : partitions, shuffle, stratégies de join, cache, lecture/écriture optimisées*
4. Ingestion batch : schémas, normalisation, gestion des dates, partitionnement temporel, layout de tables*
5. Kafka – notions pratiques : topics/partitions, clés, consommation fiable, intégration avec Spark*
6. Structured Streaming : ingestion Kafka → transformations → sink Parquet/Delta, fenêtres & watermarks (concepts), reprise après incident*
7. Qualité & Gouvernance (intro) : règles de validation, contrôles automatiques, Great Expectations (workflow & rapports)*
8. Orchestration (aperçu) : Airflow/Dagster (DAG/job), planification, idempotence, retries, notifications*
9. Industrialisation : tests (pytest), logs, configuration .env, Docker (services Spark/Kafka), traçabilité des runs*

10. Capstone : pipeline complet end-to-end (batch + streaming) avec validations, documentation et exécution reproductible

À la fin :

Vous saurez ingérer, transformer et servir des données à l’échelle avec PySpark et Kafka, produire des datasets Parquet/Delta de qualité, mettre en place des validations et des tests, et livrer une stack dockerisée avec documentation prête à l’emploi — présentable en portfolio.