Spark : Industrialisation avancée des flux de données

Lien : https://innov-systems.com/formation/spark-industrialisation-avancee-des-flux-de-donnees

DURÉE
4 jours (28h)

RÉFÉRENCE
BSI428

CATÉGORIE
Big Data : développement, langages et NOSQL

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Maîtriser les traitements avancés de données avec Apache Spark en mode batch et streaming
- Implémenter des workflows complets d’analyse et de machine learning à l’aide de Spark
- Comprendre l’utilisation de Docker pour containeriser et industrialiser les processus analytiques
- Découvrir les bonnes pratiques pour gérer le cycle de vie des modèles ML avec des outils modernes
- Appréhender l’intégration de Spark dans des architectures de traitement temps réel et big data
- Réaliser des cas pratiques couvrant des scénarios réels d’analytique et de machine learning

POUR QUI ?

- Data engineers, data scientists, analystes de données souhaitant approfondir Spark
- Développeurs et architectes désirant industrialiser des pipelines de données analytiques
- Professionnels souhaitant intégrer Spark dans des environnements cloud et conteneurisés

Programme détaillé

1 / Introduction à Spark et Docker dans l'écosystème data

- Rappels sur l'architecture et les API principales de Spark (RDD, DataFrame, Dataset)
- Concepts fondamentaux des conteneurs Docker et leur rôle dans les projets analytiques
- Mise en place d'un environnement de travail : installation et configuration Docker et Spark
- Travaux pratiques : lancement d'un cluster Spark dans des conteneurs Docker

2 / Pipeline complet de traitement analytique avec Spark

- Étapes clés : ingestion, exploration, préparation, apprentissage, déploiement
- Présentation des différentes sources de données et formats (CSV, JSON, Parquet, images, etc.)
- Les bases du traitement batch versus streaming
- Travaux pratiques : mise en place d'un pipeline d'ingestion et d'exploration sur des jeux de données variés

3 / Préparation avancée des données et feature engineering

- Techniques de nettoyage et traitement des données manquantes ou aberrantes
- Transformation et encodage des variables catégoriques, numériques et textuelles
- Création et sélection de features, réduction de dimensions et vectorisation
- Utilisation de pipelines Spark ML pour automatiser ces étapes
- Travaux pratiques : construction de pipelines de préparation complexes

4 / Machine Learning avec Spark MLlib et gestion du cycle de vie

- Introduction à MLlib : algorithmes supervisés, non supervisés, régressions, classification
- Division des jeux de données, entraînement, validation et évaluation des modèles
- Introduction à MLflow : suivi des expériences, gestion des modèles, reproductibilité
- Travaux pratiques : entraînement d'un modèle ML, tracking avec MLflow

Innov Systems

5 / Traitement temps réel et intégration avec Spark Streaming

- Concepts de streaming et micro-batch
- Connexion à Kafka, ingestion et traitement en temps réel
- Utilisation de Spark Structured Streaming pour des analyses continues
- Travaux pratiques : pipeline temps réel sur données issues de flux Kafka

6 / Industrialisation et orchestration des flux analytiques

- Conteneurisation avancée avec Docker Compose et Kubernetes
- Automatisation du déploiement et gestion des versions des pipelines
- Monitoring et alertes pour pipelines Spark en production
- Travaux pratiques : déploiement d'un pipeline analytique conteneurisé

7 / Études de cas pratiques avancées

- Recommandation produit basée sur comportement utilisateur
- Prévision de séries temporelles pour ventes ou maintenance prédictive
- Analyse sémantique et traitement de texte à grande échelle
- Intégration de Computer Vision avec Spark et bibliothèques externes (PyTorch)
- Travaux pratiques : réalisation complète d'une étude de cas au choix

8 / Approfondissements et perspectives autour du MLOps

- Principes du MLOps et rôle dans le cycle analytique
- Intégration de MLflow avec d'autres outils (CI/CD, serveurs modèles)
- Gestion des données et modèles dans des environnements collaboratifs
- Discussion ouverte : retours d'expérience, questions, conseils de mise en œuvre

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

19 au 22 Jan. 2027	Présentiel - Casablanca
16 au 19 Mars 2027	Présentiel - Casablanca
11 au 14 Mai 2027	Présentiel - Casablanca
06 au 09 Juil. 2027	Présentiel - Casablanca
07 au 10 Sep. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés