# *Spark Avancé*
## *Machine Learning et industrialisation des flux analytiques*

*Cours Pratique de 3 jours - 21h*
*Réf : SPN - Prix 2024 : 2 280€ HT*

Framework de calcul distribué, Spark permet d'effectuer des traitements et des analyses complexes en big data. Vous avez déjà utilisé Spark, nous vous proposons ici d'approfondir vos analyses avec du machine learning et de découvrir le MLOps pour le déploiement et l'industrialisation des modèles analytiques.

### OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Apprendre l'analyse avancée des données avec Spark

Effectuer des traitements de machine learning (ML) avec Spark

Comprendre Docker et son utilité dans le cadre de l'industrialisation des flux analytiques

Détailler et mettre en œuvre les étapes du cycle analytique avec Spark

Apprendre l'industrialisation du flux d'analyse

Découvrir le MLOps

### TRAVAUX PRATIQUES

Alternance de théorie et de travaux pratiques. 60?% d'exercices pour un meilleur approfondissement. Des retours d'expérience concrets.

# LE PROGRAMME

dernière mise à jour : 08/2023

### 1) Introduction

- Rappels sur l'API Spark.
- Concepts de Docker et son utilité dans les analyses de données.
- Les conteneurs Docker.

*Travaux pratiques : Prise en main de l'environnement de travail, création des conteneurs Docker.*

### 2) Le cycle analytique avec Spark

- Ingestion de données.
- Exploration.
- Préparation des données.
- Apprentissage.
- Industrialisation.

*Echanges : Présentation de cas concrets et échanges autour des différentes étapes du cycle.*

### 3) Ingestion des données.

- Le chargement de données.
- Traitements batch.
- Traitements en streaming.
- Les formats de données : images, binaires, structurés, Graph…

*Travaux pratiques : Chargement de données à partir de diverses sources.*

### PARTICIPANTS

Professionnels qui souhaitent utiliser Spark pour faire de l'analytique en mode batch ainsi qu'en temps réel.

### PRÉREQUIS

Connaissances des API Spark, notamment RDD et DataFrame. Connaissances des algorithmes d'apprentissage supervisés et non supervisés. Maîtrise d'un des langages suivants : Scala, Python.

### COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

### MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

### MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

### MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

### ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

### 4) Exploration des données
- Statistiques descriptives.
- Identifier les cas aberrants, les données vides.
- Identifier les valeurs invalides et autres anomalies.

*Travaux pratiques : Identifier les anomalies dans un jeu de données.*

### 5) Préparation et feature engineering (processus de transformation de données)
- Nettoyage des données.
- Pipelines.
- Transformer les valeurs numériques, catégoriques, binaires et texte.
- Création de nouvelles features.
- Réduction de dimensions.
- Vectorisation.

*Travaux pratiques : Préparer les données pour effectuer des analyses.*

### 6) Cycle de vie du ML avec MLflow
- Cycle de vie d'un projet de machine learning.
- Présentation de la plateforme open source MLflow.
- Les composants principaux de MLflow : Tracking, Models et Projects.
- Paramètres, métriques, balises et artefacts.

*Travaux pratiques : Création et utilisation d'un projet de machine learning.*

### 7) Machine learning
- MLlib la bibliothèque d'apprentissage automatique de Spark et les algorithmes disponibles.
- Diviser un jeu de données.
- Configurer un modèle et l'exécuter.
- Interprétation et validation de résultats d'apprentissage.
- Introduction à Spark Streaming.

*Travaux pratiques : Mise en œuvre du machine learning.*

### 8) Études de cas
- Effectuer des recommandations.
- Faire des prévisions de vente.
- Analyse sémantique.
- Computer vision avec Spark et PyTorch.
- Analyse temps réel avec Spark et Kafka.

*Etude de cas : Effectuer les différentes études de cas proposées.*

## LES DATES

| CLASSE À DISTANCE | PARIS |
|---|---|
| 2024 : 08 juil., 09 oct. | 2024 : 01 juil., 02 oct. |