

Sessions garanties disponibles

Formation PySpark

Réf. SF34593

Maîtrisez PySpark pour traiter et analyser des volumes massifs de données !

2 j Durée	1 400 € HT Inter / participant	3 700 € HT Intra (≤12 pers.)	13 Sessions à venir
--------------	-----------------------------------	---------------------------------	------------------------

Objectifs pédagogiques

- ✓ Manipuler des données massives avec Python et PySpark pour effectuer des analyses de données
- ✓ Expliquer l'architecture de PySpark et ses interactions avec les systèmes de stockage distribués tels qu'Azure Data Lake
- ✓ Configurer la connexion à des bases de données SQL Server et interagir avec elles via PySpark
- ✓ Concevoir un pipeline de transformation de données complet intégrant nettoyage, transformation et chargement en base de données

Public visé & prérequis

Public visé Data analysts, Data scientists, Ingénieurs Big Data	Prérequis Notions de base sur le langage Python ; notions sur les environnements data (modèle en étoile, Azure)
--	--

Programme détaillé

Environnement Python et prise en main

- Présentation du langage Python : positionnement, écosystème et cas d'usage en data engineering
- Installation de Python et des bibliothèques : mise en place de PySpark, pandas et des dépendances nécessaires
- Environnement virtuel : création et gestion d'un environnement isolé pour le projet
- Jupyter Notebooks : introduction à l'outil interactif pour l'exploration de données
- Bases de Python : rappel des fondamentaux du langage pour le traitement de données

Architecture et DataFrames PySpark

- DataFrames en Big Data : rôle central des DataFrames dans le traitement de données volumineuses
- Chargement des données : lecture de fichiers CSV, JSON et Parquet via PySpark
- Connexion aux systèmes de stockage : accès aux données depuis Azure Data Lake
- Filtrage, tri et agrégation : opérations de sélection et de regroupement sur les DataFrames
- Jointures entre DataFrames : croisement de plusieurs jeux de données
- Création de colonnes et transformations : enrichissement et restructuration des données
- Traitement des valeurs manquantes : détection et gestion des données incomplètes

Connexion à Azure Data Lake et SQL Server

- Architecture Azure Data Lake : présentation des couches de stockage et de leur organisation
- Lecture et écriture de fichiers : manipulation de données dans Azure Data Lake avec PySpark
- Authentification et clés d'accès : configuration sécurisée de la connexion aux ressources Azure
- Connexion SQL Server via JDBC : lecture et écriture de données dans une base relationnelle
- Exécution de requêtes SQL : sélection, transformation et extraction de données avec PySpark
- Sauvegarde des résultats : persistance des traitements dans les bases de données relationnelles

Pipeline de transformation de données

- Conception d'un pipeline complet : architecture du flux de données de bout en bout
- Chargement multi-sources : ingestion de données depuis Azure Data Lake et fichiers CSV
- Nettoyage des données : détection et correction des anomalies et incohérences

- Transformations avancées : jointures, agrégations et enrichissement des données
- Enregistrement en base de données : chargement des résultats transformés dans SQL Server
- Optimisations et bonnes pratiques : amélioration des performances et maintenabilité du pipeline

Prochaines sessions

DATE	LIEU	STATUT
03/11/2026	à distance	Sur confirmation
18/03/2027	Aix-en-Provence	Sur confirmation
18/03/2027	Grenoble	Sur confirmation
18/03/2027	Paris	Sur confirmation
18/03/2027	à distance	Sur confirmation
21/06/2027	Lyon	Sur confirmation

Pourquoi choisir Sparks Formation ?

20 ans d'expertise IT, des formateurs experts et des sessions garanties contre l'annulation.

20 ans
d'expertise

1 068+
formations

11 178
apprenants formés/an

+67
Net Promoter Score

Avis des participants

4,9/5 — 13 avis

Arnaud M. — That was really useful for me, even though I am a PDA. The trainer was very patient with me (I needed more explanation as I was the only...

Thierry R. — Formation idéale pour appréhender la technologie nosql Bon support de formation. Bonne pédagogie

Ils nous font confiance

Cette année notre filière Data Engineering & Big Data a accompagné plus de 73 entreprises dont :

SOPRA STERIA CAPGEMINI TECHNOLOGY SERVICES AMADEUS SAS GROUPAMA SUPPORTS & SERVICES AUBAY FRAIKIN FRANCE LBC FRANCE
ORANGE BUSINESS SERVICES SA ACATUS INFORMATIQUE ACENSI SANA +63 autres

Financement & certifications

Cette formation est éligible aux financements OPCO et FNE-Formation. Nos équipes vous accompagnent dans le montage de votre dossier.



Réservez votre place dès maintenant

20 ans d'expertise · 1 000+ formations · NPS +67 · réponse sous 24h ouvrées

0 805 950 800

demande@sparks-formation.com