Formation Python sur Spark avec Databricks

Objectifs : Maîtriser la librairie PySpark afin d'utiliser Apache Spark avec le langage de programmation Python sur un environnement Databricks.

Compétences visées : - Connaître l'environnement Databricks
- Se familiariser avec la librairie PySpark afin d'utiliser Apache Spark avec le langage de programmation Python
- Savoir manipuler de grands volumes de données avec Pyspark
- Mettre en oeuvre des méthodes de Machine Learning avec Pyspark

Durée : 5 jour(s) (35 heures)

Public : Développeurs, chefs de projets, data scientists, ...

Pré-requis : Pour suivre ce stage dans de bonnes conditions, il est recommandé d'avoir suivi en amont la formation Python – Bases et introduction aux librairies scientifiques

Méthode pédagogique : Pédagogie active mêlant exposés, exercices et applications pratiques. La formation s'effectue sur un environnement Databricks.

Modalités d'évaluation : Un formulaire d'auto-évaluation proposé en amont de la formation nous permettra d'évaluer votre niveau et de recueillir vos attentes. Ce même formulaire soumis en aval de la formation fournira une appréciation de votre progression.
Des exercices pratiques seront proposés à la fin de chaque séquence pédagogique pour l'évaluation des acquis.
En fin de formation, vous serez amené(e) à renseigner un questionnaire d'évaluation à chaud.
Une attestation de formation vous sera adressée à l'issue de la session.
Trois mois après votre formation, vous recevrez par email un formulaire d'évaluation à froid sur l'utilisation des acquis de la formation.

Accessibilité : Vous souhaitez suivre notre formation Python sur Spark avec Databricks et êtes en situation de handicap ? Merci de nous contacter afin que nous puissions envisager les adaptations nécessaires et vous garantir de bonnes conditions d'apprentissage

Tarifs :
- Présentiel : 3250 € HT
- Distanciel : 3000 € HT
(-10% pour 2 inscrits, -20% dès 3 inscrits)

Option(s) :
- Forfait déjeuners : 100 € HT

Nos prochaines sessions

Distance
du 8 au 12 décembre 2025

Lyon
du 19 au 23 mai 2025
du 27 au 31 octobre 2025

Paris
du 23 au 27 juin 2025
du 24 au 28 novembre 2025

Toulouse
du 13 au 17 octobre 2025

Programme :

– Présentation de Databricks

Cette introduction permet de vous initier à l'environnement Databricks et ses outils

- Historique
- Différence entre l'utilisation Administrateur et Utilisateur
- Comment mettre en place un projet sur Databricks/AWS
- Comment créer un cluster de calcul avec Databricks/AWS
- Gestion des notebooks, des utilisateurs et des ressources

– Introduction à Spark

Spark est un environnement de travail distribué qui permet d'effectuer des calculs sur des gros volumes de données

- Rappels sur le Big Data
- Présentation de Spark: Spark RDD, Spark SQL, Spark MLlib, Spark GraphX
- Configurer un Spark Context et une Spark Session
- Gestion de la mémoire sous Spark

– Présentation de PySpark et l'API Pyspark RDD

Pyspark est l'API Python de Spark

- Présentation de Pyspark et fonctionnement avec les Java Virtual Machines
- Présentation de l'API Pyspark RDD et manipulation de données non structurées
- Mise en pratique avec des fichiers textes (comptage de mots, nettoyage d'un fichier texte structuré) et avec des opérations d'agrégation sur PairRDDs

– Utilisation de l’API Pyspark SQL

L’API Pyspark SQL permet de manipuler des données structurées sous format de Dataframes avec du Python et du SQL

- Présentation de l’API Pyspark SQL
- Lecture de fichiers csv, json, parquet et sauvegarde de fichier
- Nettoyage et manipulation de données
- Groupby et agrégation
- Jointure de tables
- Lien entre SQL et Python
- Manipulation d’objets Row, et Window
- Manipulation de dates
- Utilisation de User Defined Function et de Pandas User Defined functions
- Présentation de Pyspark Pandas
- Nombreuses mises en pratique sur des jeux de données

– Introduction au Machine Learning

Rappels des fondamentaux du Machine Learning

- Qu’est-ce que le Machine Learning? Apprentissages supervisé et non supervisé
- Compromis Biais Variance
- Modèles Linéaires
- Modèles Non Linéaires
- Modèles ensemblistes
- Modèles de clustering
- Métriques et évaluation des performances

– Machine Learning avec PySpark

Les librairies associées à Pyspark MLlib permettent de faire tourner des modèles de Machine Learning dans un environnement de calcul distribué

- Différence entre MLlib SQL et MLlib RDD
- Les transformations de processing sur les données et notions de Pipeline
- Mise en pratique avec des modèles de Machine Learning sur des problématiques de clustering, de classification (données numériques et texte)
- Evaluation des performances avec Pyspark MLlib
- Utilisation de la librairie Xgboost sous un environnement Spark
- Utilisation de Pandas UDF pour faire tourner des modèles scikit-learn ou tensorflow en inférence

Date de dernière modification : 5 novembre 2024