Formation Python pour le Big Data

Objectifs : Utiliser le langage Python pour manipuler et visualiser de grands ensembles de données (big data) en exploitant ses nombreuses librairies scientifiques

Compétences visées : - Connaître les problématiques du Big Data
- Connaître les différentes librairies Python permettant de manipuler le Big Data
- Savoir manipuler de grands volumes de données
- Avoir des notions sur l'architecture Big Data

Durée : 5 jour(s) (35 heures)

Public : Architectes, développeurs, data scientists, chefs de projet, ...

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 dans le logiciel Python.

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 Formation par ville 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 30 juin au 4 juil. 2025
du 3 au 7 novembre 2025

Lyon
du 9 au 13 décembre 2024
du 14 au 18 avril 2025
du 8 au 12 décembre 2025

Paris
du 2 au 6 décembre 2024
du 16 au 20 juin 2025
du 1 au 5 décembre 2025

Toulouse
du 24 au 28 mars 2025
du 29 sept. au 3 oct. 2025

Programme :

- Concepts du Big Data

Cette introduction permet de vous initier à la problématique du Big Data

- Volume, Vitesse, Véracité
- Map Reduce
- Architecture Big Data et Data Lake
- Big Data et Cloud computing
- Les outils du Big Data

- Introduction à la librairie Dask

Dask est une librairie qui permet de faire du calcul distribué sur plusieurs cœurs ou plusieurs machines avec la possibilité d'utiliser un scheduler. Dask peut donc accélérer le calcul sur de larges volumes de données.

- Présentation de Dask
- Exemple de calculs distribués
- Dask et Numpy: comparaison de performances
- Dask et Pandas

- Introduction à la librairie Xarray

Xarray est une librairie Python qui s'appuie sur Numpy et permet de manipuler de larges volumes de données. Cette librairie est particulièrement efficace pour des fichiers netCDF et peut s'utiliser de concert avec Dask

- Présentation de Xarray
- Exemples d'utilisation de Xarray
- Mise en pratique avec un fichier netCDF

- Introduction à la librairie Vaex

Vaex est une librairie qui ressemble beaucoup à pandas mais qui fait des calculs à la volée sans gaspiller l'usage de la RAM. On peut dès lors traiter des données qui ont près de 1 milliard de lignes à la seconde.

- Présentation de Vaex
- Prise en main de Vaex avec des exemples
- Comparaison entre Vaex et Pandas
- Visualisation des données avec Vaex

- Introduction à Spark

Spark est un outil permettant le passage à l'échelle pour la gestion des données et le calcul distribué. Bien que géré par Apache, Spark est en Open Source et peut s'utiliser avec plusieurs langages dont Python

- Présentation de Spark
- Architecture Apache Spark
- Autres outils associés à Spark (Yarn, Mesos)
- Resilient Distributed Dataset (RDD)
- Présentation et Installation de PySpark

- Introduction à PySpark

Vous verrez grâce une mise en pratique sur une journée la prise en main de PySpark, comment lire et gérer des données, comment appliquer des fonctions sur les données et comment appliquer une réduction de dimension

- Visualisation des données massives avec Holoviews

Manipuler des gros volumes de données n'est pas toujours suffisants, on veut pouvoir aussi les visualiser. La librairie Holoviews permet aussi bien de transformer des données massives que de les visualiser.

- Présentation et prise en main d'Holoviews
- Interactivité avec Holoviews

Date de dernière modification : 5 novembre 2024