Python pour l'ingénierie de données avec Pandas, Polars et écosystèmes volumineux

Lien : https://innov-systems.com/formation/python-pour-lingenierie-de-donnees-avec-pandas-polars-et-ecosystemes-volumineux

DURÉE
5 jours (35h)

RÉFÉRENCE
BSI422

CATÉGORIE
Big Data : développement, langages et NOSQL

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Maîtriser Pandas pour l'ingénierie et l'analyse de données courantes
- Savoir choisir et utiliser des alternatives performantes (Polars, Dask, DuckDB, Spark) selon le volume et l'architecture
- Nettoyer, profiler et valider des jeux de données réels avec des outils industriels
- Exploiter CPU et GPU et comprendre les compromis entre mémoire, vitesse et facilité d'utilisation
- Produire des visualisations et prototypes de dashboards pour communiquer les résultats

POUR QUI ?

- Développeurs, data engineers, data analysts et scientifiques de données ayant des bases en Python (variables, fonctions, structures de contrôle)
- Personnes souhaitant manipuler des jeux de données allant de quelques Mo à plusieurs dizaines de Go, sur poste local ou en environnement distribué

Programme détaillé

1 / Fondamentaux robustes et bonnes pratiques avec Pandas

- Panorama pratique de Pandas et des alternatives — cas d'usage, limites, bonnes pratiques de performance
- Lecture/écriture de données (CSV, Parquet, Excel, JSON, SQL) avec gestion optimisée des types et chunks
- Manipulations avancées : filtrage, groupby, merges, fenêtres temporelles, rolling windows
- Optimisation mémoire et profiling basique (downcasting, vectorisation)
- Atelier pratique : pipeline ETL simple (lecture → nettoyage → agrégation)

2 / Nettoyage, profilage, et validation robustes

- Profilage exploratoire (missingness maps, résumé statistique, visualisation rapide)
- Nettoyage : traitement des valeurs manquantes, outliers, standardisation de formats
- Validation par schémas et règles (contrats de données)
- Mise en place de tests automatisés sur flux ETL
- Atelier : pipeline complet de profilage + nettoyage + validation

3 / Alternatives performantes et bases de données embarquées

- Polars vs Pandas : lazy execution, streaming, migration de code
- DuckDB : requêtes SQL sur fichiers locaux Parquet/CSV, intégration Python
- Dask et Spark : principes, cas d'usage, parallélisation locale et distribuée
- Comparaison pratique sur un même jeu de données
- Atelier : choix de la bonne bibliothèque pour un cas réel et implémentation

4 / CPU, GPU et traitement sur plusieurs machines

- Différences CPU/GPU : architecture, transferts, cas pertinents
- Utilisation de bibliothèques optimisées GPU-friendly

Inov Systems

- Traitement multi-machines : chunking, repartitioning, réduction du shuffle
- Atelier comparatif : même pipeline exécuté en mode CPU, GPU et partitionné

5 / Visualisation, dashboards et mini-projet final

- Visualisation selon le volume de données : matplotlib/seaborn (petits), Datashader/Vaex (grands)
- Bonnes pratiques : agrégations, échantillonnage, interactivité
- Prototypage rapide : Streamlit, Dash, Panel
- Mini-projet intégrateur : lecture, nettoyage, transformation, visualisation et dashboard d'un dataset complet
- Présentation des projets, feedback, remise des livrables et ressources pour aller plus loin

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

18 au 22 Jan. 2027	Présentiel - Casablanca
15 au 19 Mars 2027	Présentiel - Casablanca
10 au 14 Mai 2027	Présentiel - Casablanca
05 au 09 Juil. 2027	Présentiel - Casablanca
30 Août au 03 Sep. 2027	Présentiel - Casablanca

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

Inov Systems

Réservation & Renseignements

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

Innov Systems — Tous droits réservés

Innov Systems