Data Engineer, Bootcamp (13 semaines)

by DataScientest

Cours pratique - 57j - 399h00 - Réf. 4II

Prix : 7490 € H.T.

Devenez expert en analyse de données avec ce parcours proposé par notre partenaire DataScientest. Un Data Engineer a pour mission de concevoir des outils et solutions qui vont permettre de traiter et d'analyser de grands volumes de données. Cette formation certifiante se déroule à distance dans un format hybride mêlant temps d'échanges synchrones avec un formateur expert, exercices pratiques et modules E-learning. Basée sur la pédagogie Learning By Doing, vous réaliserez un projet fil rouge en équipe afin de mettre en pratique vos connaissances. Lors de votre inscription, vous serez rattaché à l'une des promotions Datascientest. A l'issue de cette formation, vous obtiendrez un co-certificat « Data Engineer » des Mines Paris - PSL Executive et de DataScientest ainsi que les blocs de compétence 2 et 3 de la certification RNCP « Data Engineer ». Contactez-nous dès maintenant pour connaître les prochaines dates !

Objectifs pédagogiques

À l'issue de la formation, le participant sera en mesure de :

- Élaborer une architecture technique de gestion de données.
- Déployer une solution d'analyse de données massives intégrant l'intelligence artificielle.

Public concerné

Personnes ayant une appétence pour la programmation et la manipulation des données.

Prérequis

Un diplôme ou un titre de niveau bac+3 et des connaissances en Python, SQL, Linux.
Pour les candidats ne présentant pas le niveau de qualification requis, une dérogation est possible sur dossier.

PARTICIPANTS

Personnes ayant une appétence pour la programmation et la manipulation des données.

PRÉREQUIS

Un diplôme ou un titre de niveau bac+3 et des connaissances en Python, SQL, Linux.
Pour les candidats ne présentant pas le niveau de qualification requis, une dérogation est possible sur dossier.

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.

Certification

Pour clôturer la formation, l'équipe pédagogique évaluera le projet fil rouge de l'apprenant à l'aide d'un rapport écrit et d'une soutenance à distance. La validation des compétences développées au cours de la formation Data Engineer vous permettra d'obtenir : • Un co-certificat « Data Engineer » des Mines Paris - PSL Executive et de DataScientest • Les blocs de compétence 2 et 3 de la certification RNCP de niveau 7 "Data Engineer" enregistrée au RNCP sous le n°RNCP38919.

Méthodes et moyens pédagogiques

Activités digitales

Cours et exercices en ligne, masterclass collective, séances de questions/réponses, classes de soutien, accompagnement par mail, projet fil rouge, coaching carrière individualisé, social learning.

Tutorat

Un formateur expert accompagne l'apprenant tout au long de sa formation. Il échange régulièrement avec lui sur son projet fil rouge et l'accompagne lors de points de mentorat (individuel). Plusieurs formateurs animent également les différentes masterclass (classes collectives) et répondent aux questions des apprenants à tout moment depuis un forum dédié. En complément, de nombreuses séances de questions-réponses peuvent être organisées pour aider les apprenants.

Pédagogie et pratique

Lors de l'inscription, l'apprenant est affecté à une promotion (dates à définir lors de l'inscription) et reçoit son calendrier de formation. Le parcours de formation est découpé en « Sprint » de plusieurs semaines sur une thématique dédiée. Chaque semaine l'apprenant est convié à un temps d'échange avec le formateur qui se présente sous la forme de masterclass (classe collective) ou de points de mentorat (individuel). Pendant 80% du temps, l'apprenant travaille en autonomie sur la plateforme d'enseignement. Tous les modules intègrent des exercices pratiques permettant de mettre en œuvre les concepts développés en cours. L'apprenant doit également travailler en binôme ou trinôme sur un projet fil rouge tout au long de la formation. Cela lui permettra de développer et faire reconnaître ses compétences. En complément, des événements et ateliers thématiques sont régulièrement proposés pour permettre aux apprenants de découvrir les dernières innovations en matière de Data Science. Afin de suivre efficacement la formation, nous estimons le temps travail nécessaire entre 35 et 40 heures par semaine.

Programme de la formation

1 Prochaines dates de sessions

- Octobre 2025 : Début au 07/10/25
- Novembre 2025 : Début au 04/11/25
- Décembre 2025 : Début au 02/12/25

2 Programmation

- Python : variables, types, opérateurs, boucles, fonctions, classes, modules.
- Python : multithreading et multiprocessing sur Python, fonction asynchrone, bibliothèque MyPy.
- Web Scraping : BeautifulSoup, navigation sur un document HTML et identification des données.

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 formations pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque formation 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 participant 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

Pour toute question ou besoin relatif à l'accessibilité, vous pouvez joindre notre équipe PSH par e-mail à l'adresse psh-accueil@orsys.fr.

3 Outils avancés

- Git : introduction, dépôt Git, branche, tag et merge.
- GitHub : introduction à la plateforme, fork, pull request, issues, pull and push, Github Actions.
- Système Linux et Script Bash : systèmes Linux, utilisation d'un terminal, scripts Bash.

4 Big Data Variété

- SQL : bases de données relationnelles, langage SQL, approfondissement et application.
- ElasticSearch : moteur de recherche, index, Mapping, Ingest node, Text Analyzer.
- MongoDB : présentation, requêtes MongoDB.
- Neo4j : données orientées graph, requête Cypher, chargement de données, client Python pour Neo4J.
- Hbase : bases de données orientées colonne, modification des données par Python et Happybase.

5 Batch & streaming

- PySpark : calcul distribué, APIs RDD et Dataframe, processing de données distribuées, Machine Learning distribué.
- Kafka : architecture et avantages, gestion des paramétrages, paramétrages de Consumers.
- Streaming avec Spark : traitement de données temps réel, mini-batch streaming, Structured Streaming, pipeline.

6 Entreposage pratique des données

- Snowflake : Data Warehousing avec une sécurité robuste, analyse de données SQL pour le cloud, optimisation plateforme.
- Data Warehousing avec DBT (ELT) : transformations, datasets de haute qualité, automatisation de l'exécution.

7 Cloud AWS

- AWS Solution Architect : bonnes pratiques, conception d'architectures, amélioration continue et automatisation.
- AWS Solution Architect : présentation du cloud AWS, les services clés de la plate-forme AWS.

8 Machine Learning

- Statistiques : variables numériques, variables catégorielles, relations entre les variables.
- Data Visualisation : différents types de graphiques avec Matplotlib, création d'applications Dash.
- Machine Learning : pré-traitement, algorithmes de Machine Learning (régression, classification, clustering).
- ML Flow : l'architecture MLFlow, MLFlow Tracking, MLFlow Projects, MLFlow Models, MLFlow Registry, cycle de vie.

9 DevOps - Virtualisation

- APIs : architectures micro services, méthodes HTTP, librairies FastAPI et Flask, spécification OpenAPI, gestion API.
- Docker : concept de conteneurisation, images et des conteneurs, communication, persistance, Dockerhub, docker-compose.
- Sécurisation des API : clés API (API Keys), authentification HTTP Basique, JSON Web Token et HTTPS.
- Kubernetes : déployer et gérer des conteneurs, initialisation et architecture, API avec Kubernetes.

10 CI/CD et Monitoring

- Airflow : concept d'orchestration, graphe orienté acycliques ou DAG, opérateurs, gestion des tâches, monitoring.
- Tests unitaires avec Python : tests unitaires avec Pytest, tests d'intégration, avantages des tests, intégration.
- GitLab : installation, initialisation, ajout et suppression, Git Blame, Tag, statut de dépôt, gestion des conflits.
- Prometheus & Grafana : utilité du monitoring, Prometheus Query Language, Dashboard avec Grafana, intégration.