# *Data Engineer, parcours intensif de 11 semaines tutorat inclus*

*Formation en ligne - 400h*
*Réf : 4II - Prix 2024 : 5 990€ HT*

ORSYS et DataScientest, leader dans le domaine de la data science en France, proposent un parcours de formation complet pour exercer le métier de data engineer. Le data engineer ou expert big data est le « monsieur » IT des équipes data, il gère l'architecture de la donnée et mets en production les modèles conçus par le data scientist. Ce parcours en ligne vous apprendra les fondamentaux de Linux et des scripts bash, du langage Python et des bases de données relationnelles et NoSQL. Il aborde également l'usage des technologies autour de la data science (statistiques, machine learning, etc.), des systèmes de gestion de versions comme Git ou GitHub ainsi que l'automatisation et le déploiement d'application.

## OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Maîtriser les fondamentaux de Linux et des scripts bash

Connaître les fondamentaux du langage Python

Utiliser les bases de données relationnelles et NoSQL

Utiliser les technologies autour de la data science (statistiques, machine learning...)

Mettre en place l'automatisation et le déploiement d'applications

Appréhender les systèmes de gestion de versions comme Git ou GitHub

## PÉDAGOGIE ET PRATIQUES

Formation digitale basée sur une pédagogie active et conçue avec des experts en data science. Une combinaison de théorie, de démonstrations, de mises en pratique, de partages d'expériences et de bonnes pratiques. Un test de positionnement, un accompagnement sur mesure dès le début du parcours, un projet fil rouge et des séquences pédagogiques de courte durée permettent de renforcer l'apprentissage et d'évaluer l'apprenant tout au long de sa formation. En cas de besoin technique, une cellule support est disponible en ligne 5 jours sur 7 de 9h à 18h30. À l'issue de son parcours en ligne, l'apprenant reçoit un certificat délivré par MINES ParisTech | PSL Executive Education, ce qui lui permet de bénéficier de la reconnaissance d'un acteur de référence dans le domaine.

## ACTIVITÉS DIGITALES

Test de positionnement sous forme de QCM d'une heure, séance d'introduction à la plateforme à distance, classes collectives, classe de soutien sur mesure, simulation et codage en direct, exercices, fiches de synthèse, projet fil rouge dédié, social learning, échanges avec data scientists.

## PARTICIPANTS

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

## PRÉREQUIS

Avoir un niveau bac +3 en mathématiques ou un niveau bac+5 en sciences (ingénieur, mathématique, statistique, économétrie…). Une formation en ingénierie ou informatique est conseillée.

## COMPÉTENCES DU FORMATEUR

Les experts qui ont conçu la formation et qui accompagnent les apprenants dans le cadre d'un tutorat sont des spécialistes des sujets traités. 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. 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

La progression de l'apprenant est évaluée tout au long de sa formation au moyen de QCM, d'exercices pratiques, de tests ou d'échanges pédagogiques.
Sa satisfaction est aussi évaluée à l'issue de sa formation grâce à un questionnaire.

## MOYENS PÉDAGOGIQUES ET TECHNIQUES

Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : documentation et support de cours, exercices pratiques d'application et corrigés des exercices, études de cas ou présentation de cas réels.
ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques. Une attestation de fin de formation est fournie si l'apprenant a bien suivi la totalité de la formation.

## MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

## ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

# LE PROGRAMME

dernière mise à jour : 03/2023

### 1) Systèmes Linux & Python

- Présentation des systèmes Linux.
- Prise en main et utilisation d'un terminal.
- Mise en place de scripts bash.
- Maîtrise des variables et des types.
- Présentation des divers opérateurs et de leurs applications.
- Introduction au concept de boucles et aux structures de contrôle.
- Définition d'une fonction sur Python et de leurs applications.
- Initiation aux classes et modules.
- Préparation de la mise en place, du paramétrage et de l'enchainement de décorateurs.
- Différenciation et implémentation du multithreading et du multiprocessing sur Python.
- Application d'une fonction asynchrone sur Python.
- Introduction aux annotations et utilisation de la bibliothèque mypy.

### 2) SQL et MongoDB

- Introduction aux bases de données relationnelles.
- Présentation de SQLAlchemy et applications.
- Initiation aux bases du langage SQL.
- Approfondissement de SQL et de ses applications.
- Introduction aux bases de données NoSQL (base de données orientée document, colonne, graphe).
- Présentation de MongoDB.
- Familiarisation avec la syntaxe des requêtes MongoDB.

### 3) Elasticsearch et Neo4j

- Description d'un moteur de recherche.
- Présentation d'un index et mode d'emploi.
- Mise au point d'un mapping.
- Découverte des différentes opérations.
- Prétraitement des données avec ingest node.
- Extraction des données avec les text analyzers.
- Introduction aux bases de données orientées graphe.
- Mise en place d'un premier graphe.
- Initiation au langage de requête Cypher.
- Chargement de données dans Neo4J.
- Utilisation d'un client Python pour Neo4J.

### 4) Statistiques et machine learning

- Exploration des variables numériques.
- Exploration des variables catégorielles.
- Étude des relations entre les variables.
- Prétraitement de données.
- Sélection et optimisation d'un algorithme de machine learning.
- Définition et application d'un algorithme de régression.
- Définition et application d'un algorithme de classification.
- Développement d'algorithmes de clustering.
- Introduction au PCA (Principal component analysis, analyse en composantes principales).

### 5) Dataviz avec matplotlib

- Découverte de graphes : en barres (barplot), nuages de points (scatter plot), histogrammes, camembert (pie chart), ...

### 6) Hadoop/Hive et HBase/Spark

- Fonctionnement de Hadoop.
- Installation et configuration de Hadoop.
- Traitement et stockage des données avec HDFS.
- Présentation de MapReduce.
- Utilisation de Hadoop streaming pour exécuter un fichier MapReduce.
- Mise en place d'entrepôts de données.
- Présentation du fonctionnement de Hive.
- Présentation des bases de données orientées colonne.
- Association de Hadoop (HDFS) et de HBase. Requêtes de données.
- Modification des données par Python et HBase.
- Distinction entre Spark et Hadoop.
- Introduction au calcul distribué avec Spark.
- Présentation des API, RDD et dataframe de Spark.
- Pipeline de processing de données distribuées avec PySpark.
- Machine Learning distribué avec Spark MLlib.

### 7) Git, GitHub et quality assurance

- Introduction au système de gestion de version Git.
- Initialisation d'un dépôt Git.
- Présentation et approfondissement des concepts Git : branches, tag, merge.
- Découverte de la plateforme GitHub pour le travail collaboratif sur Git.
- Présentation des fonctionnalités majeures de GitHub : fork, pull, request, issues.
- Partager ses modifications avec pull et push.
- Participation à l'amélioration de projets publics (open source).
- Présentation des principaux workflows Git.
- Mise en place de tests unitaires avec pytest.
- Introduction aux Tests d'intégration et leurs fonctions.
- Présentation des avantages des tests : gain de temps, lisibilité, qualité et amélioration de code.

### 8) Architecture de streaming Kafka et Spark Streaming

- Gestion de flux de données en temps réel.
- Conception d'une architecture big data hybride (batch et temps réel).
- Mise en place d'une architecture lambda.
- Présentation de la plateforme de streaming distribuée Kafka : architecture, avantages.
- Gestion des paramétrages de producers : clef de partitionnement.
- Maîtrise des paramètres de consumers : consumer group.
- Prise en main de Spark Streaming pour le traitement de données en temps réel.
- Présentation du MiniBatch streaming nécessaire pour le fonctionnement de Spark Streaming.

### 9) API

- Introduction aux API et découverte des architectures microservices.
- Présentation des différentes méthodes HTTP et de leurs fonctions.
- Utilisation des librairies FastAPI et Flask pour développer des API RESTful.
- Documentation d'une API avec la spécification OpenAPI.
- Gestion des erreurs et des performances d'une API.
- Découverte d’Airflow : orchestration, graphe orienté acycliques ou DAG (directed acyclic graphs) et opérateurs.
- Gestion de tâches par le biais d'opérateurs spécifiques.
- Monitoring des DAG (directed acyclic graphs) via l'interface graphique d'Airflow.

### 10) Docker et Kubernetes

- Présentation de la conteneurisation et de son utilité par rapport à la virtualisation.
- Initiation au fonctionnement de Docker.
- Manipulation des images et des conteneurs.
- Communication avec les conteneurs.
- Persistance des données grâce aux volumes.
- Création d'une image Docker via un Dockerfile.
- Partage des images sur le Docker Hub.
- Utilisation de Docker Compose.
- Déploiement et gestion des conteneurs.

## NOS POINTS FORTS

- Séquences de courte durée
- Activités digitales variées
- Accès illimité pendant 1 an ou pendant la durée du parcours
- Tutorat personnalisé inclus ou en option
- Accès multi-device (smartphone, tablette ou ordinateur)