

# Devenir Data Engineer

**97 jours | 679 heures**

Tous nos cursus sont réalisables, au niveau national, en formation distancielle.



En partenariat avec :



Document mis à jour le 21/03/2023

Date de formation : Du 27 mars au 1<sup>er</sup> septembre 2023  
Lieu(x) de formation : Villeneuve d'Ascq  
Réunion d'Information : Le 13 mars 2023  
Date limite d'envoi de candidature : Le 10 mars 2023

## OBJECTIFS DE FORMATION

Le/La Data Engineer ~~dé~~init, développe, met en place et maintient les outils et infrastructures adéquats à l'analyse de la donnée par les équipes de Data Science. Il/Elle veille à créer une solution permettant le traitement de volumes importants de données tout en garantissant la sécurité de celles-ci.

Il/Elle représente le premier maillon de la chaîne de traitement de données.

Cette fonction exige un vrai sens de l'organisation et de la méthode ainsi qu'un sens aigu du travail en équipe.

Au terme de cette formation, vous serez en mesure :

- d'apporter une expertise en Big Data permettant la manipulation des données
- de concevoir les plateformes permettant de traiter des volumes importants données
- de mettre en place des bases de données
- de veiller à ce que les pipelines de données soient sécurisés et clairs pour être analysés et transformés

## PRÉREQUIS

- Vous êtes titulaire d'un BAC +4 minimum scientifique ou informatique
- Vous avez un excellent relationnel et êtes capable de vous adapter à différents interlocuteurs : clients, interlocuteurs techniques, etc.
- Vous êtes bon(ne) communicant(e), appréciez le travail en équipe et faites preuve d'excellentes capacités rédactionnelles
- La maîtrise de l'anglais professionnel ainsi que des notions en Python, SQL et/ou Linux seront fortement appréciées

## PUBLIC CONCERNÉ ET FINANCEMENT

Cette formation ~~fin~~ancée à 100% est à destination des demandeur(se)s d'emploi inscrit(e)s à Pôle emploi.

# COMPÉTENCES ET TECHNOLOGIES ABORDÉES

Module	Objectifs du module	Durée (Jours)	Durée (Heures)
Présentation du cursus, des plateformes pédagogiques	Ice Breaking, tour de table - Avoir une description du cursus et des choix d'orientation - Comprendre l'utilisation des différentes plateformes d'apprentissage du cursus -	0,5	3,5
Apprendre à apprendre	Comprendre comment fonctionne le cerveau pour mémoriser plus efficacement - Mettre en place des outils et méthodes d'apprentissage et de mémorisation - Comprendre l'utilisation des différentes plateformes d'apprentissage du cursus - Avoir une description du cursus et des choix d'orientation.	0,5	3,5
Adopter la posture gagnante du consultant en ESN	Maîtriser le contexte des ESN - Maîtriser les différents types de prestations - Comprendre les cycles de ventes - Comprendre le métier de consultant - Adapter votre posture - Identifier vos axes de progressions.	1	7
Algorithmique et programmation structurée orientée Python	Disposer des connaissances nécessaires à l'apprentissage d'un langage de développement - Connaître les structures de base de la programmation (boucles, conditions) - Savoir quelles sont les grands paradigmes de programmation (procédural, objet) - Comprendre la notion d'objet et les concepts associés - Identifier les apports de la modélisation UML - Disposer d'un premier point de vue sur les approches Python - Découvrir les variables et le typage des données.	3	21
Python et Python Orienté Objet dans Visual Studio Code	Identifier les usages courants du langage - Installer la prise en charge de Python dans Visual Studio - Mettre en pratique le scripting en Python - Structurer votre code en fonction, classes et modules - Utiliser des modules existants - Décrire la programmation réseau avec Python - Expérimenter la programmation objet en Python.	7	49
Mises en pratique et Validation des acquis - Python	Etudes de cas - Quiz - Exercices - Projet groupé	2	14
PEP 8 : Apprendre les bonnes pratiques pour coder en Python	La PEP 8, C'est quoi ? - Encodage - L'indentation - Code layout - Import - Les espaces - Commenter son code en anglais	1	7
Linux - les fondamentaux	Distinguer les logiciels libres et l'Open Source - Identifier les principes fondamentaux du système d'exploitation - Utiliser interactivement le Shell et connaître les commandes essentielles - Gérer les fichiers et les dossiers - Editer un fichier - Reconnaître les métacaractères et les expressions régulières.	4	28
Linux - Programmation Shell Bash	Décrire les notions de bases de Linux, de manière approfondie - Enrichir votre corpus de commandes - Développer des scripts Shell.	4	28
Mises en pratique et Validation des acquis - Infrastructure Linux	Etudes de cas - Quiz - Exercices - Projet groupé	2	14
Langage SQL : Fondamentaux	Décrire les principaux concepts des SGBDR (Système de Gestion des Bases de Données Relationnelles) et d'algèbre relationnelle utilisés dans le langage SQL - Interroger une base de données avec la clause SQL SELECT - Utiliser les commandes SQL de mise à jour des données - Identifier les commandes SQL de début et fin de transaction BEGIN, COMMIT et ROLLBACK - Présenter les concepts de gestion des privilèges systèmes et objets avec les commandes SQL GRANT et REVOKE - Créer, modifier et supprimer certaines catégories d'objets (table, index, vues...) avec CREATE, ALTER et DROP.	3	21
Agilité et Pilotage de projets	Gestion de projet : prédictive vs agile - Pourquoi l'agilité ? - Apprendre les pratiques agiles - Maîtriser la méthode agile Scrum .	3	21
Big Data - Enjeux et perspectives	Identifier l'ensemble des enjeux et facteurs à prendre en compte pour réussir l'intégration du Big Data dans la vision large du SI Evaluer et sélectionner les outils appropriés dans le cadre d'un plan de mise en œuvre du Big Data.	2	14

Big Data - Concevoir et piloter un projet	Bien identifier les challenges dans la mise en place d'un projet Big Data Mesurer l'impact politique et stratégique sur l'entreprise et la pertinence d'un reporting régulier Mettre en place une bonne gestion des conflits et de la conduite du changement.	2	14
Hadoop - Introduction à la plateforme et à l'écosystème	Installer un environnement Hadoop - Utiliser les distributions Hortonworks, MapR et Cloudera - Réaliser des algorithmes de calcul distribué avec MapReduce - Comprendre les concepts principaux autour du stockage et du traitement des données dans Hadoop - Appréhender l'écosystèmes basés sur Hadoop (HBase, les outils de traitement Pig, Hive et Spark)	1	7
Hadoop - Développement	Présenter les principes du Framework Hadoop - Utiliser la technologie MapReduce pour paralléliser des calculs sur des volumes importants de données - Identifier les commandes shell courantes pour HDFS.	2	14
PySpark - Traitement des données	Décrire le principe de fonctionnement de Spark - Utiliser l'API PySpark pour interagir avec Spark en Python - Mettre en œuvre les méthodes de Machine Learning avec la librairie MLlib de Spark - Traiter les flux de données avec Spark Streaming - Manipuler les données avec Spark SQL.	3	21
Mises en pratique et Validation des acquis - Big Data	Etudes de cas - Quiz - Exercices - Projet groupé	1	7
TECHNIQUE RECHERCHE D'EMPLOI	Atelier cv et entretien d'embauche	1	7
Data Visualisation - Rendre visible l'invisible	Analyser, trier, traiter des données brutes pour les présenter sous forme de graphiques, de cartes ou d'organigrammes attrayants - Simplifier la lisibilité et accroître la compréhension.	1	7
Power BI - Les fondamentaux à l'expertise	Décrire le cycle de création d'un rapport Power BI - Vous connecter à des sources de données - Transformer, nettoyer et combiner des sources - Structurer un modèle de données - Créer des indicateurs - Afficher les indicateurs dans des rapports PBI - Partager des rapports - Mettre en place une solution de Business Intelligence avec les outils avancés Power BI - Les fonctions avancées du DAX - Partager en ligne vos tableaux de bords et rapports - Utiliser des visualisations interactives.	5	35
Qlik Sense - Du designer au développeur	Décrire les avantages et les fonctionnalités de Qlik Sense - Naviguer dans l'application Qlik Sense - Gérer les feuilles dans une application Qlik Sense - Manipuler les objets de feuille pour choisir ou restituer des données - Organiser les objets sur la feuille - Créer un graphique pour mettre en valeur les données - Charger simplement des données dans Qlik Sense - Exploiter l'outil de développement - Vous connecter aux sources de données - Nettoyer, manipuler et transformer des données - Réaliser des modélisations associatives et des optimisations - Utiliser des fichiers QVD - Réaliser des calculs avancés (Mêlres, agrégations, totaux et sous-totaux...) - Utiliser les éléments principaux.	4	28
Data Science - Les fondamentaux	Identifier les enjeux de la Data Science et de l'IA - Schématiser le cycle d'un projet Data Science - Appliquer les meilleures pratiques en matière de nettoyage et de préparation de vos données avant l'analyse - Modéliser un problème de Data Science.	2	14
Mises en pratique et Validation des acquis - Data Visualisation et Data Science	Etudes de cas - Quiz - Exercices - Projet groupé	3	21
Fondamentaux Devops	Décrire les principes de la démarche DevOps - Démontrer l'impact de la démarche DevOps dans les infrastructures et notamment l'Infrastructure as Code - Reconnaître l'impact de la démarche DevOps, des conteneurs et du CaaS - Entamer une démarche vers une organisation DevOps.	2	14
Gérer le versionning avec GIT et appliquer le TDD avec PyTest	Découverte de la plateforme Github pour le travail collaboratif sur Git - Décrire les principes d'un gestionnaire de versions distribué - Identifier par la pratique, la philosophie de Git et ses apports - Mettre en œuvre les principes fondamentaux et les bonnes pratiques du TDD - Installer le framework PyTest - Gérer les tests logiciels avec PyTest sur un projet Agile.	3	21

Mises en pratique et Validation des acquis - CI/ CD	Etudes de cas - Quiz - Exercices - Projet groupé	1	7
Airflow - Utiliser l'orchestrateur de workflows	Découvrir les fonctionnalités d'Apache Airflow. - Savoir concevoir, planifier et monitorer ses workflows. - Connaître les mécanismes avancés de l'outil. - Savoir comment scaler Airflow afin de gérer un nombre illimité de tâches à exécuter. - Connaître les best practices et limitations de Airflow.	3	21
TECHNIQUE RECHERCHE D'EMPLOI	Atelier cv et entretien d'embauche	1	7
Docker pour Linux - Déploiement de conteneurs virtuels	Identifier les caractéristiques d'un conteneur Linux - Installer et utiliser Docker - Décrire la création des images Docker et "Dockerfile" - Interagir avec le Docker Hub et registry privés - Lister les notions réseaux de Docker (networks, links) - Expliquer la gestion des données avec Docker (volumes) - Utiliser Docker Swarm, Docker Compose et Docker Machine - Intégrer Docker au sein de vos projets - Déployer Docker en production.	3	21
Installer Kubernetes avec Minikube	Fonctionnalités de Minikube - Installation de Minikube - Démarrer, utiliser et supprimer les minikube localement - Gérer et interagir avec votre cluster - Mise en réseau - Volumes persistants - Dossiers hôtes montés - Gestion des Add-ons - Utilisation de Minikube avec un proxy HTTP.	1	7
Kubernetes - Orchestrer ses conteneurs	Décrire les principes de l'orchestration de conteneurs Docker - Créer et mettre en œuvre Kubernetes - Orchestrer des conteneurs Docker - Déployer des applications.	3	21
Mises en pratique et Validation des acquis - Kubernetes avec Minikube	Etudes de cas - Quiz - Exercices - Projet groupé	1	7
Google Cloud Platform - Core infrastructure - Les fondamentaux	Démontrer l'utilité et les atouts des produits et services Google Cloud Platform - Interagir avec les services Google Cloud Platform - Décrire les manières dont les clients utilisent Google Cloud Platform - Choisir et utiliser des environnements de déploiement applicatif sur Google Cloud Platform : Google App Engine, Google Kubernetes et Google Compute Engine - Gérer les options de stockage Google Cloud Platform : Google Cloud Storage, Google Cloud Filestore, Google Cloud SQL, Google Cloud BigTable et Google Cloud Data Store - Décrire les principes de base d'utilisation de BigQuery, l'entrepôt de données géré par Google et destiné à l'analyse - Présenter les principes de base d'utilisation de Cloud Deployment Manager, l'outil de Google permettant de créer et gérer des ressources Cloud à l'aide de modèles - Identifier les principes de base d'utilisation de Google Stackdriver, le système de surveillance, de journalisation et de diagnostic de Google.	1	7
Google Kubernetes Engine - Architecting	Décrire le déploiement et la gestion d'applications conteneurisées sur Google Kubernetes Engine (GKE) et les autres services fournis par Google Cloud Platform (GCP) - Explorer et déployer des éléments de solution, notamment des composants d'infrastructure tels que des pods, des conteneurs, des déploiements et des services (ainsi que des réseaux et des services d'application) - Déployer des solutions, notamment la gestion de la sécurité et des accès, la gestion des ressources et la surveillance des ressources - Décrire le fonctionnement des conteneurs de logiciels - Justifier les choix GCP pour les services de stockage managés - Surveiller les applications exécutées dans Kubernetes Engine - Illustrer l'architecture de Kubernetes - Concevoir l'architecture de GCP - Expliquer le fonctionnement de la mise en réseau des pods dans Kubernetes Engine - Créer et gérer des clusters de Kubernetes Engine à l'aide de la console GCP et des commandes gcloud / kubectl - Lancer, annuler et exposer des jobs dans Kubernetes - Gérer le contrôle d'accès à l'aide de Kubernetes RBAC et Google Cloud IAM - Gérer les stratégies de sécurité des pods et des réseaux - Utiliser des "secrets" et des "ConfigMaps" pour isoler les informations d'identification de sécurité et les artefacts de configuration.	3	21
Big Data and Machine Learning Fundamentals on Google Cloud Platform	Connaissance des produits et services de GCP notamment ceux liés aux traitements des données et au machine learning - Connaissance des produits et services fondamentaux concernant le calcul et le stockage - Connaissance de Cloud SQL et de Dataproc - Connaissance de Datalab et BigQuery - Connaissance de TensorFlow et des APIs machine learning (ML) - Connaissance de Pub/Sub et de Dataflow.	1	7

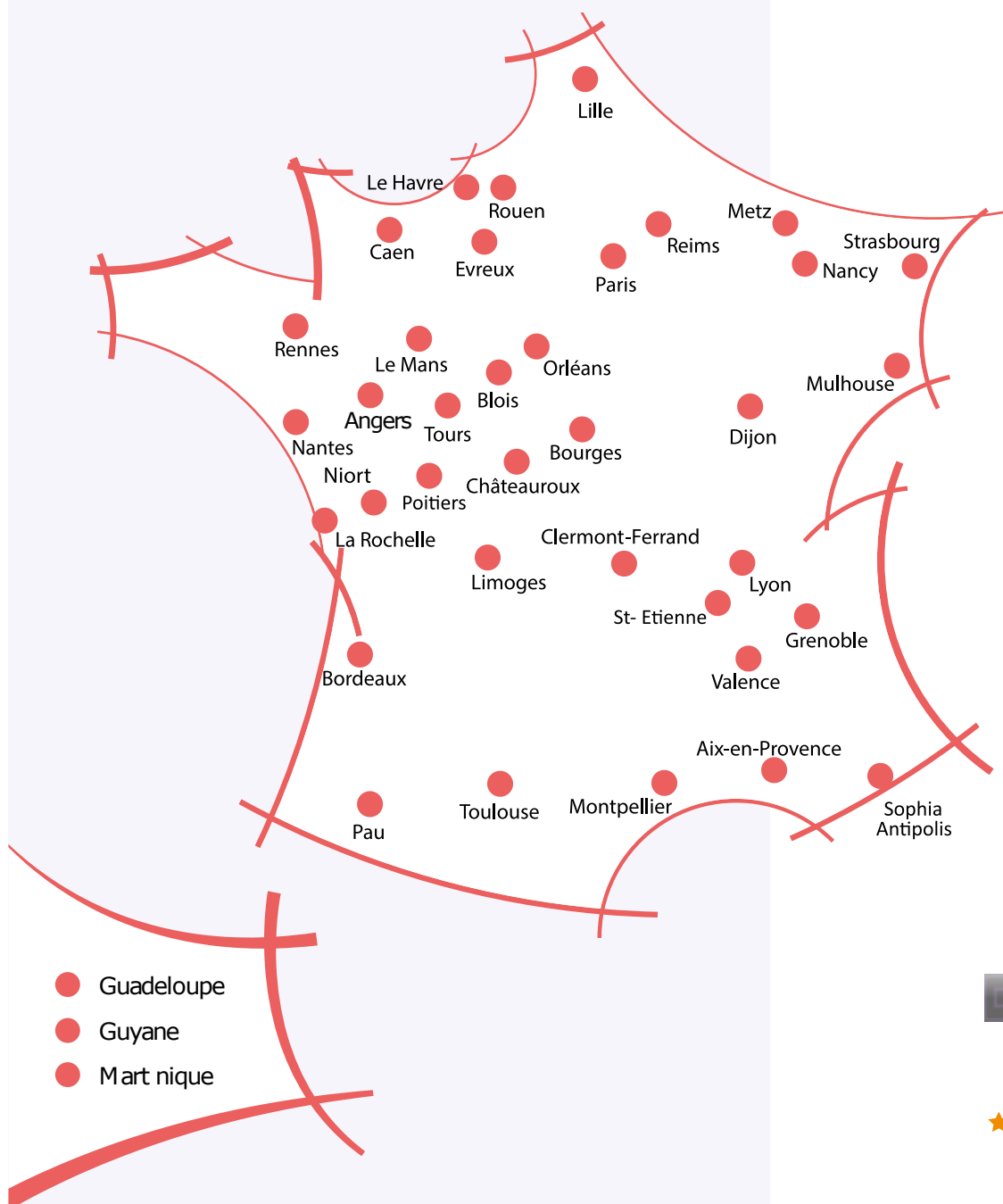
Langage SQL avancé avec BigQuery	Décrire de manière approfondie les techniques du langage SQL afin de mieux tirer profit des possibilités du langage dans la consultation et la manipulation des données.	3	21
Data Engineering on Google Cloud Platform	Concevoir et développer des systèmes de traitement des données sur Google Cloud - Traiter des données par lot ou par flux en mettant en œuvre des pipelines de données d'autoscaling sur Dataflow - Orchestrer le travail entre les services GCP avec Cloud Composer / Apache Airflow - Obtenir des insights métier à partir d'ensembles de données extrêmement volumineux à l'aide de BigQuery - Exploiter des données non structurées à l'aide de Spark et des interfaces de programmation de ML sur Dataproc - Obtenir des insights immédiats à partir de flux de données - Découvrir les API de machine learning (ML) et BigQuery ML, et apprendre à utiliser Cloud AutoML	4	28
Analyse et visualisation des données dans Looker	Définir Looker et les fonctionnalités qu'il fournit pour travailler avec des données - Expliquer les quatre concepts analytiques de base dans Looker (dimensions, mesures, filtres, pivots) - Utilisez des dimensions, des mesures, des filtres et des pivots pour analyser et visualiser les données - Créez instantanément des métriques avancées avec des calculs de table - Créer des tableaux de bord pour combiner et partager des visualisations - Utilisez des dossiers et des tableaux dans Looker pour organiser le contenu pour la navigabilité et la découvrabilité.	1	7
Sauvegarde quotidienne d'un GCP Cloud Storage Bucket	Présentation générale des services de backup fournis par GCP - Création du Bucket de sauvegarde - Création de la Cloud Function - Création du Cloud Scheduler - Les différentes méthodes de restauration.	1	7
Mises en pratique et Validation des acquis - Google Cloud Platform, BigQuery, Cloud Composer, Airflow, etc.	Etudes de cas - Quiz - Exercices - Projet groupé	7	49

**97 jours (679h)**

## LES PLUS DE M2IFORMATION

- Formation animée par des formateurs certifiés
- Mise à disposition du matériel (ordinateur)
- Une expertise de 40 ans en formation informatique
- 70% de la formation dédiés à la pratique

**INSCRIVEZ-VOUS**



ORGANISME DE FORMATION  
**INCONTOURNABLE**



**m2iformat on.fr**