L'AKADÉMIE DU DIGITAL · PROGRAMME DE FORMATION Certification Qualiopi en cours de préparation

Data Engineer

FOR-DEV-M6-A · Développement, DevOps & Data

Concevoir, construire et exploiter les pipelines et infrastructures de données

DURÉE
35 heures — 5 jours

TARIF
2?450 € HT

PUBLIC
Écoles & centres de formation

DÉLAI D'ACCÈS
Accès sous 2 jours ouvrés après validation du besoin et des prérequis.

Tarif établi sur la base de 35h à 70 € HT/h. Parcours dédié aux écoles et centres de formation, établi sur devis ; tarif indicatif et négociable selon le volume et le contexte.

Le module alterne séquences théoriques et ateliers pratiques tout au long de chaque journée. Il couvre la modélisation des données, le SQL avancé et les entrepôts, les pipelines ETL/ELT et leur orchestration, le big data et le traitement distribué, les architectures modernes et la gouvernance. Le fil rouge des ateliers est SQL/Python/Airflow/Spark/Kafka, complété par un panorama étendu de l'écosystème.

Objectifs pédagogiques

- Comprendre le rôle du data engineer et l'architecture d'une plateforme de données
- Concevoir et modéliser des bases de données (relationnelles et NoSQL) et un entrepôt de données
- Construire des pipelines de données (ETL/ELT) et les orchestrer
- Comprendre les enjeux du big data et du traitement distribué (batch/streaming)
- Connaître les architectures modernes (data warehouse, data lake, lakehouse) et le cloud data
- Mettre en place qualité, gouvernance, sécurité et supervision des données

Public concerné

- Étudiants et apprenants des écoles et centres de formation partenaires
- Développeurs ou data analysts souhaitant s'orienter vers l'ingénierie des données
- Profils techniques visant des fonctions data engineering

Prérequis

Des bases de programmation (Python recommandé) et de SQL sont requises, ainsi qu'une aisance avec la ligne de commande. Des notions de bases de données et de data science facilitent la progression. Ces éléments sont validés par un formulaire en ligne et / ou un entretien préalable.

- Maîtriser les bases de la programmation (Python de préférence) et de SQL.

- Être à l'aise avec la ligne de commande et la manipulation de fichiers.
- Avoir des notions de bases de données et de data (modules Full-Stack et Data Scientist conseillés).
- Disposer d'une autonomie technique et d'une rigueur d'organisation.

Programme

Jour 1 — Le métier & modélisation des données

Métier & architecture data	1h15	Data engineer vs data scientist ; architecture d'une plateforme ; panorama de l'écosystème (bases, ETL, distribué, cloud). Exposé interactif, schéma d'architecture, panorama d'outils.	Questions orales ; QCM sur l'écosystème.
Analyser une architecture	2h00	Identifier les composants d'une architecture data. TP guidé, étude d'architecture, grille d'analyse.	Analyse d'architecture validée.
Modélisation des données	1h15	Relationnel (PostgreSQL, MySQL), NoSQL (MongoDB, Cassandra, Redis), normalisation, modélisation décisionnelle (étoile/flocon). Exposé, panorama des bases, exemples de schémas.	QCM sur la modélisation.
Modéliser un schéma	2h00	Concevoir un schéma de données adapté à un cas. TP, outil de modélisation, accompagnement.	Schéma de données validé.

Jour 2 — SQL avancé & entrepôts de données

SQL avancé	1h15	Jointures complexes, fonctions de fenêtrage, optimisation, index. Exposé, exemples de requêtes, plans d'exécution.	Questions orales ; QCM.
Requêtes SQL avancées	2h15	Écrire des requêtes avancées sur un jeu de données. TP guidé, base de données, requêtes progressives.	Requêtes avancées validées.
Entrepôts de données	1h00	Data warehouse, modélisation dimensionnelle, panorama (BigQuery, Snowflake, Redshift). Exposé, schémas dimensionnels, panorama cloud.	QCM sur les entrepôts.
Mini-entrepôt	2h30	Construire et alimenter un mini-entrepôt de données. TP, modélisation en étoile, chargement, requêtes.	Mini-entrepôt validé.

Jour 3 — Pipelines ETL/ELT & orchestration

Ingestion & ETL/ELT	1h15	Ingestion, ETL vs ELT, transformation, idempotence ; panorama (dbt, Talend, Fivetran, Airbyte, NiFi). Exposé, schémas de pipelines, panorama d'outils.	Questions orales ; QCM.
Pipeline ETL	2h15	Construire un pipeline ETL (extraction ? transformation ? chargement). TP guidé Python, sources et cible, accompagnement.	Pipeline ETL fonctionnel.
Orchestration de pipelines	1h00	Airflow (fil rouge), planification, dépendances ; alternatives Dagster, Prefect. Exposé, notion de DAG, panorama des orchestrateurs.	QCM sur l'orchestration.
Orchestrer avec un DAG	2h30	Orchestrer un pipeline via un DAG (Airflow). TP guidé Airflow, planification, gestion des dépendances.	DAG orchestré validé.

Jour 4 — Big data & traitement distribué

Enjeux du big data	1h15	Volume, vitesse, variété ; traitement distribué (Spark) ; héritage Hadoop/MapReduce, alternatives Flink, Dask. Exposé, schémas distribués, panorama des moteurs.	Questions orales ; QCM.
Traitement distribué (Spark)	2h15	Traiter un gros volume en distribué (initiation Spark). TP guidé Spark, jeu de données volumineux, accompagnement.	Traitement distribué validé.
Batch vs streaming	1h00	Batch vs temps réel ; Kafka (fil rouge), alternatives Pulsar, Kinesis, RabbitMQ. Exposé, schéma streaming, panorama des bus de messages.	QCM sur batch/streaming.
Traitement batch & démo streaming	2h30	Mettre en place un traitement batch ; démonstration streaming. TP batch, démonstration Kafka, revue collective.	Traitement batch validé.

Jour 5 — Architectures modernes, gouvernance & projet

Architectures modernes & cloud	1h00	Data warehouse / data lake / lakehouse ; formats ouverts (Delta Lake, Iceberg, Hudi) ; cloud (BigQuery, Snowflake, Redshift, Databricks, Synapse, Dataflow). Exposé, schémas d'architectures, panorama cloud.	QCM sur les architectures.
Explorer une architecture cloud	2h00	Explorer une architecture cloud data (initiation). TP guidé, environnement cloud d'initiation, accompagnement.	Exploration cloud validée.
Qualité & gouvernance	1h00	Qualité, gouvernance, sécurité, RGPD, supervision des données. Exposé, bonnes pratiques, exemples de supervision.	QCM sur la gouvernance.
Pipeline data de bout en bout & restitution	2h30	Concevoir et présenter un pipeline data complet. TP projet, présentations, revue collective, QCM final, enquête de satisfaction.	Évaluation sommative : pipeline data + QCM final.

Modalités et moyens pédagogiques

Formation en présentiel dans nos locaux ou dans vos locaux, en groupe restreint. Alternance d’apports théoriques, d’ateliers pratiques et de mises en situation.

Matériel requis :

16 Go de Ram recommandés

- Les stagiaires doivent disposer d'un PC avec le Wifi et 16 Go de RAM recommandés.
- Effectif adapté au format école / centre de formation.
- Parcours dédié aux écoles et centres de formation, établi sur devis.

Modalités d’évaluation

Évaluation des acquis en cours et en fin de formation, attestation de fin de formation remise à chaque stagiaire.

Accessibilité

Formation accessible aux personnes en situation de handicap. Contactez-nous avant l’inscription pour étudier les aménagements nécessaires.

Financement

Devis ou convention de formation sur demande. La certification Qualiopi est en cours de préparation : l’éligibilité aux financements OPCO sera ouverte après son obtention.

Audit de certification prévu au T1 2027. Aucune éligibilité OPCO avant l’obtention.