Orchestration de pipelines avec Apache Airflow

Référence : **BDDAPA003** Durée : **2 jours (14 heures)** Certification : **Aucune**

CONNAISSANCES PRÉALABLE

- Avoirs des connaissances en Python et en ETL

PROFIL DES STAGIAIRES

- Data engineers, DevOps

OBJECTIFS

- Créer et planifier des DAGs
- Superviser et gérer les workflows
- Automatiser des tâches répétitives

CERTIFICATION PRÉPARÉE

- Aucune

MÉTHODES PÉDAGOGIQUES

- Mise à dispositon d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique numérique pendant le stage
- La formation est constituée d'apports théoriques interactif, démonstrations en direct, travaux dirigés, ateliers collaboratif, mises en situation professionnelles, projet fil rouge, études de cas réels, ateliers d'industrialisation et d'automatisation, ateliers de supervision et de diagnostic
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement par demi-journée de formation

FORMATEUR

- Consultant-formateur expert Apache Airflow

MÉTHODES D'ÉVALUATION DES ACQUIS

- Validation des prérequis envoyée en amont de la formation. En l'absence de réponse d'un ou plusieurs participants, un temps sera consacré en ouverture de session pour prendre connaissance du positionnement de chaque stagiaire sur les objectifs pédagogiques évalués
- Validation progressive des compétences tout au long de la formation via exercices pratiques, quiz techniques, études de cas, création de DAGs, résolution d'incidents simulés et restitution des travaux réalisés dans le cadre du projet fil rouge
- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation des compétences acquises envoyée au stagiaire
- Attestation de fin de stage adressée avec la facture

ACCESSIBILITÉ DE LA FORMATION

- Si vous avez besoin d'un aménagement pour suivre votre formation dans de bonnes conditions, nous vous invitons à contacter notre référent handicap au 01 71 19 70 30 ou à l'adresse referent.handicap@edugroupe.com. Un échange en amont nous permettra d'étudier avec vous les adaptations possibles.

CONTENU DU COURS

1. JOUR 1 - Concevoir et orchestrer des workflows avec Airflow (7h)

2. Comprendre l'architecture Apache Airflow (2h)

- Présentation d'Apache Airflow
- Cas d'usage de l'orchestration
- Architecture Airflow
- Scheduler
- Executor
- Metadata Database
- Webserver
- Concepts fondamentaux : DAG, Task, Operator
- Cycle d'exécution d'un workflow
- 💡 *Mises en pratique : Découverte d'un environnement Airflow ; Exploration de l'interface Web ; Analyse de DAGs existants*

3. Créer et développer des DAGs (3h)

- Syntaxe des DAGs Python
- Operators standards
- PythonOperator
- EmailOperator
- Dépendances entre tâches
- Gestion des paramètres
- Variables et connexions Airflow
- 💡 *Mises en pratique : Création d'un DAG simple ; Construction d'un workflow multi-étapes ; Gestion des dépendances et paramètres*

4. Planifier et automatiser les traitements (2h)

- Scheduling
- Cron expressions
- Gestion des exécutions
- Catchup
- Backfill
- Retry policies
- Gestion des erreurs
- Bonnes pratiques d'automatisation
- 💡 *Mises en pratique : Planification de workflows ; Configuration de stratégies de reprise ; Tests d'exécution automatisée*

5. JOUR 2 - Superviser, administrer et industrialiser les pipelines (7h)

6. Superviser et administrer les workflows (3h)

- Interface d'administration Airflow
- Monitoring des DAGs
- Journaux d'exécution
- Gestion des erreurs
- Alertes et notifications
- Analyse des dépendances
- Outils de diagnostic
- 💡 *Mises en pratique : Analyse de workflows en échec ; Exploitation des logs ; Résolution d'incidents simulés*

7. Intégrer Airflow dans une architecture Data (2h)

- Connexions aux bases de données
- Intégration avec des APIs
- Déclenchement de traitements Spark
- Intégration avec Kafka
- Exécution de scripts Python
- Gestion des flux ETL et ELT
- 💡 *Mises en pratique : Construction d'un pipeline multi-systèmes ; Orchestration de traitements de données ; Automatisation d'un flux complet*

8. Projet fil rouge : industrialisation d'un pipeline de données avec Airflow (2h)

- Analyse du besoin
- Conception du DAG
- Planification
- Gestion des erreurs
- Supervision
- Documentation
- Industrialisation
- 💡 *Mises en pratique : Création d'un pipeline complet de bout en bout ; Orchestration de plusieurs traitements ; Mise en place de mécanismes de surveillance ; Présentation des résultats et débriefing collectif*