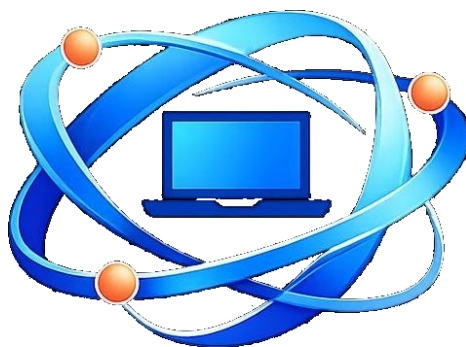




MFE-IT

Référence : MFE-AWSDEA

Formation AWS Data Engineer – Pipelines de données sur AWS

Construire et opérationnaliser des pipelines de données sur AWS

Durée : 3 jours | Volume horaire : 21 h

Distanciel · Sessions garanties dès 1 inscrit · 60 % de pratique

DESCRIPTION

La formation AWS Data Engineer s'attaque à ce qui se passe entre la donnée brute et le tableau de bord : l'ingestion, la transformation, le stockage, et surtout la partie dont personne ne parle — faire en sorte que le pipeline tourne encore dans six mois, qu'on sache d'où vient chaque chiffre et combien il a coûté.

OBJECTIFS PÉDAGOGIQUES

À l'issue de cette formation, les participants seront capables de :

- Concevoir une architecture de données sur AWS : lac de données, entrepôt et modèle lakehouse.
- Ingérer des données par lots et en flux continu (Kinesis, DMS, DataSync).
- Transformer des données à grande échelle avec AWS Glue et Amazon EMR.
- Choisir et optimiser les formats et le partitionnement du stockage sur S3 (Parquet, Iceberg).
- Modéliser, charger et optimiser un entrepôt Amazon Redshift.
- Orchestrer des pipelines et y intégrer des contrôles de qualité de données.
- Gouverner les accès aux données avec Lake Formation et maîtriser le coût des traitements.

PRÉREQUIS

- Une pratique du SQL est indispensable : vous devez savoir écrire une jointure et une agrégation sans hésiter. Des notions de Python facilitent nettement les modules Glue et EMR, sans être formellement exigées.
- Une première expérience d'AWS (S3, IAM, EC2) est attendue. Le niveau de la certification AWS Cloud Practitioner constitue une base suffisante.
- Parce que chaque participant est unique, un entretien personnalisé avec notre expert nous permet de concevoir une formation parfaitement alignée avec ses objectifs.

PUBLIC VISÉ

- Ingénieurs data et développeurs ETL construisant des pipelines sur AWS.
- Analystes et développeurs BI amenés à alimenter eux-mêmes leurs sources de données.
- Administrateurs de bases de données évoluant vers les architectures de lac de données.
- Architectes et responsables techniques préparant la certification Data Engineer – Associate.

PROGRAMME DÉTAILLÉ

La formation alterne apports théoriques et pratique (environ 60 % du temps). Les modules s'articulent autour d'exercices concrets basés sur des cas d'usage métier réels.

Module 1 — Architecture de données sur AWS

Trois modèles et ce qui les sépare réellement : lac de données sur S3, entrepôt Redshift, et le lakehouse qui prétend réconcilier les deux. Panorama honnête des services et de leurs recouvrements — AWS en propose souvent trois pour le même besoin. Zones du lac (brut, nettoyé, exploitable). Le catalogue de données AWS Glue comme couche de métadonnées commune. Critères de décision : volume, latence, profil des utilisateurs, coût.

Module 2 — Ingestion par lots et en flux

Faire entrer la donnée, quelle que soit sa forme. Par lots : AWS DMS pour répliquer des bases relationnelles, y compris en capture des changements (CDC), DataSync et Transfer Family pour les fichiers. En flux : Amazon Kinesis Data Streams (partitions, clé de partitionnement, rétention) et Data Firehose pour la livraison gérée vers S3 ou Redshift. Amazon MSK quand Kafka est déjà en place. Gestion des schémas et de leur évolution. Lab : brancher une source transactionnelle sur le lac, en flux.

Module 3 — Stockage, formats et partitionnement

Là se joue une grande partie de la performance et de la facture. Formats : pourquoi Parquet et ORC écrasent CSV et JSON dès que le volume monte. Partitionnement : le bon découpage, et le piège des petits fichiers qui étranglent les traitements. Compression. Formats de table transactionnels avec Apache Iceberg : mises à jour, suppressions, voyage dans le temps, évolution de schéma. Classes de stockage S3 et cycles de vie. Lab : mesurer l'effet d'un repartitionnement sur le temps et le coût d'une même requête.

Module 4 — Transformation avec Glue et EMR

AWS Glue : crawlers et catalogue, jobs Spark, Glue Studio pour le visuel, signets pour le traitement incrémental. Amazon EMR quand il faut la main sur Spark : dimensionnement du cluster, EMR Serverless, instances Spot pour les traitements tolérants. Bases de l'optimisation Spark : partitions, mise en cache, et le déséquilibre de données qui fait durer un job trois heures au lieu de dix minutes. Choix entre Glue et EMR selon la charge et l'équipe.

Module 5 — Entrepôt Amazon Redshift

Architecture MPP et ce qu'elle implique dans la modélisation. Clés de distribution et de tri : les deux décisions qui déterminent les performances, et qu'on regrette longtemps si on les rate. Chargement massif avec COPY. Redshift Spectrum pour interroger S3 sans charger. Redshift Serverless. Vues matérialisées, compression, maintenance (VACUUM, ANALYZE). Diagnostic d'une requête lente : lire un plan d'exécution plutôt que deviner.

Module 6 — Requêtage, orchestration et qualité

Amazon Athena : SQL directement sur S3, facturation à la donnée scannée — d'où l'enjeu du partitionnement vu plus tôt. Orchestration : Step Functions, workflows Glue, Amazon MWAA (Airflow géré) — critères de choix. Reprise sur erreur, idempotence, retraitement d'une plage de dates. Qualité des données : règles Glue Data Quality, détection d'anomalies, que faire d'un lot invalide. Lab : orchestrer un pipeline complet et le faire échouer proprement.

Module 7 — Gouvernance, sécurité et coût

Qui a le droit de voir quoi : AWS Lake Formation, permissions au niveau table, colonne et ligne, partage entre comptes. Chiffrement KMS et gestion des données sensibles (Amazon Macie). Traçabilité et lignage. Le volet qu'on découvre trop tard : le coût. D'où vient la facture d'un pipeline (données scannées, heures de calcul, stockage), et les leviers réels — formats, partitionnement, Spot, dimensionnement. Atelier de synthèse : auditer un pipeline lent et coûteux, produire un plan d'optimisation chiffré.

LES PLUS DE CETTE FORMATION

- Construit une chaîne data complète, de l'ingestion en flux jusqu'à la restitution, sur des volumes réalistes.
- Suit le référentiel de la certification AWS Certified Data Engineer – Associate (DEA-C01), qui remplace la Data Analytics Specialty.
- Traite frontalement le coût et la qualité des données, là où la plupart des projets data déraillent.
- Se déroule en groupe de 1 à 3 participants, ce qui permet de travailler sur vos propres jeux de données.

MODALITÉS PÉDAGOGIQUES

Format et déroulement

La formation est dispensée à distance via une classe virtuelle interactive. Elle peut également être réalisée sur site, avec un contenu personnalisé selon les besoins de votre projet professionnel. La répartition théorie/pratique est d'environ 40 %/60 %.

Format ultra-personnalisé MFE-IT

Chaque session accueille de 1 à 3 participants, garantissant un accompagnement très individualisé. Un entretien préalable permet d'adapter le contenu au profil de chacun. Les sessions inter-entreprises sont garanties dès 1 seul inscrit (pas de risque de report sauf cas de force majeure).

Évaluation des compétences

Tout au long de la formation, le formateur évalue la progression des participants par des QCM, des mises en situation et des travaux pratiques. À la fin, une attestation de validation des acquis est remise à chaque participant.

Suivi post-formation

Pendant un mois après la formation, chaque participant peut contacter les formateurs MFE-IT pour toute question sur la mise en œuvre des acquis. Une réponse est apportée par e-mail ou téléphone sous 48 heures ouvrées.

Accessibilité

MFE-IT s'engage à accueillir les personnes en situation de handicap. Contact : contact@mfe-it.com.

INFORMATIONS PRATIQUES

Ressources formateur

- Démonstrations structurées alignées sur le programme détaillé
- Énoncés d'exercices et corrigés tout au long de la formation
- Un environnement technique prêt à l'emploi pour les ateliers pratiques
- Validation des acquis par le formateur à la fin de chaque atelier
- Documents de référence numériques

Certification et validation

À l'issue de la formation, une attestation est envoyée par e-mail précisant les objectifs, la nature, la durée et les résultats de l'évaluation. Un certificat de réalisation peut également être fourni sur demande.

Bénéfices pour les participants

- Se former depuis son lieu de travail ou son domicile, sans déplacement
- Bénéficier d'un formateur-consultant expert du sujet
- Profiter d'un format ultra-personnalisé (1 à 3 participants)
- Poursuivre la formation même en cas d'imprévu

Bénéfices pour l'organisation

- Optimiser le budget formation en réduisant déplacements et hébergements
- Offrir une formation de qualité à tous les collaborateurs, où qu'ils soient
- Réduire le temps d'absence lié aux déplacements
- Accompagner la montée en compétences des équipes dans tous les contextes