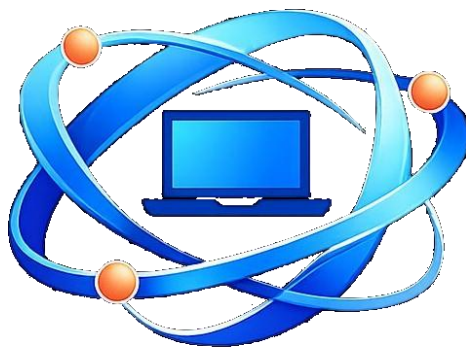




MFE-IT

Référence : DP-750

Formation Azure Databricks (DP-750) : Ingénierie des données

Concevoir et exploiter des pipelines de données sur Azure Databricks

Durée : 4 jours | Nombre d'heures : 24 h

Distanciel · Sessions garanties dès 1 inscrit · 60 % de travaux pratiques

DESCRIPTION

Cette formation intensive vous outille pour concevoir, déployer et exploiter des solutions d'ingénierie des données sur Azure Databricks. Alignée sur le référentiel de l'examen Microsoft DP-750 (parcours Azure Data Engineer Associate), elle couvre l'ensemble de la chaîne : configuration de l'environnement, gouvernance avec Unity Catalog, ingestion et transformation des données, pipelines Lakeflow et optimisation des charges Spark. Elle prolonge naturellement les fondamentaux de la formation Microsoft Azure DP-900 et s'appuie sur un socle Azure que vous pouvez consolider avec la formation Microsoft Azure AZ-104.

Sur 4 jours à forte dominante pratique, vous construisez pas à pas un pipeline de données moderne : création de catalogues et de tables Delta, ingestion en streaming avec Auto Loader, transformation et contrôle qualité, orchestration via Lakeflow Jobs, puis surveillance et optimisation. Chaque notion est immédiatement mise en œuvre en atelier sur un espace Databricks. Les sessions se déroulent en petit groupe de 1 à 3 participants.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation, vous serez capable de :

- Configurer un espace de travail Azure Databricks (compute, Photon, runtime) et organiser les objets Unity Catalog.
- Sécuriser et gouverner les données : privilèges, contrôle d'accès aux tables, colonnes et lignes, lineage et audit.
- Ingérer des données en batch et en streaming avec Lakeflow Connect, Auto Loader et Spark Structured Streaming.
- Nettoyer, transformer et fiabiliser les données avec des contraintes de qualité et la gestion du schema drift.
- Concevoir, planifier et déployer des pipelines Lakeflow (Jobs, pipelines déclaratifs, Asset Bundles, CI/CD Git).
- Surveiller, dépanner et optimiser les charges Spark et les tables Delta (OPTIMIZE, VACUUM, Azure Monitor).

PRÉREQUIS

Parce que chaque participant est unique, un entretien préalable ajuste le contenu. Une aisance de base avec SQL et Python ainsi que des notions de Git et de Data Lake sont recommandées. Une première expérience d'Azure est un plus, sans être indispensable.

PUBLIC VISÉ

Cette formation vise les data engineers, les développeurs et architectes data, ainsi que les analystes souhaitant industrialiser leurs traitements sur Azure Databricks, et toute personne préparant la certification DP-750 (Azure Data Engineer Associate).

PROGRAMME DÉTAILLÉ

La formation alterne apports théoriques et travaux pratiques (environ 60 % du temps). Les modules sont construits autour d'exercices pratiques inspirés de cas d'usage métier concrets.

Module 1 — Configurer un environnement Azure Databricks

- Choisir et configurer le compute (job, serverless, entrepôt SQL, classique, partagé) : autoscaling, Photon, runtime, pooling.
- Installer des bibliothèques et gérer les permissions d'accès au compute.
- Créer et organiser les objets Unity Catalog : catalogues, schémas, volumes, tables, vues et vues matérialisées.
- Appliquer des conventions de nommage et implémenter des catalogues externes (foreign catalogs).

Module 2 — Sécuriser et gouverner Unity Catalog

- Accorder des privilèges aux principaux (utilisateurs, groupes, principaux de service).
- Mettre en place le contrôle d'accès au niveau des tables, des colonnes et des lignes (row-level security).
- Accéder aux secrets Azure Key Vault et authentifier l'accès via des identités managées.
- Gouverner : ABAC par tags, masques de colonnes, rétention, traçabilité (lineage), journaux d'audit et Delta Sharing.

Module 3 — Modéliser et ingérer les données

- Concevoir la logique d'ingestion et choisir les outils (Lakeflow Connect, notebooks, Azure Data Factory).
- Choisir les formats de tables (Delta, Parquet, Iceberg) et les stratégies de partitionnement et de clustering (liquid clustering, Z-ordering).
- Ingérer en batch et en streaming : Auto Loader, Spark Structured Streaming, Azure Event Hubs, CDC, COPY INTO.
- Gérer les dimensions à évolution lente (SCD) et les tables temporelles.

Module 4 — Nettoyer, transformer et fiabiliser les données

- Profiler les données, choisir les types de colonnes, traiter doublons et valeurs manquantes.
- Transformer : filtres, agrégations, jointures, unions, pivots et dénormalisation.
- Charger via les opérations merge, insert et append.
- Contraintes de qualité : validations, application de schéma, gestion du schema drift, attentes des pipelines (expectations).

Module 5 — Concevoir et déployer les pipelines

- Concevoir l'ordre des opérations et la gestion des erreurs des pipelines.
- Créer des pipelines avec notebooks et Lakeflow Spark Declarative Pipelines.
- Implémenter les Lakeflow Jobs : déclencheurs, planification, alertes, redémarrages automatiques.
- Cycle de vie : bonnes pratiques Git, tests, Databricks Asset Bundles, déploiement via CLI et API REST.

Module 6 — Surveiller, dépanner et optimiser

- Surveiller la consommation des clusters pour la performance et le coût.
- Dépanner les Lakeflow Jobs et les jobs Spark (goulots, redémarrage, tuning).
- Résoudre caching, skew, spilling et shuffle via le Spark UI et le profil de requête.
- Optimiser les tables Delta (OPTIMIZE, VACUUM) et le monitoring via Azure Monitor / Log Analytics.
-

LES PLUS DE LA FORMATION

- Une pédagogie orientée cas réels : chaque compétence est ancrée dans un atelier Databricks concret.
- Une prise en main de Unity Catalog, Lakeflow et Spark, préparant à la certification DP-750.
- Un petit groupe (1 à 3 participants) et un accompagnement de 30 jours après la session.
-

MÉTHODES PÉDAGOGIQUES

Format et déroulement

Formation en distanciel, en groupe de 1 à 3 participants, à raison de 6 h par jour. Le présentiel peut se dérouler directement sur votre site, et le distanciel peut être enregistré à la demande. Environ 60 % du temps est consacré aux travaux pratiques sur un espace de travail Azure Databricks.

Format ultra-personnalisé MFE-IT

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

Évaluation des acquis

Tout au long de la formation, le formateur évalue la progression des participants à l'aide de questions à choix multiples, de mises en situation et de travaux pratiques. À l'issue, une attestation de réussite est remise à chaque participant.

Accompagnement post-formation

Pendant un mois après la formation, chaque participant peut contacter les formateurs MFE-IT pour toute question relative à la mise en œuvre des connaissances acquises. Une réponse est apportée par e-mail ou par 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 du 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 au format numérique

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. Une attestation de fin de formation peut également être fournie 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

- Proposer une formation de qualité à tous les collaborateurs, où qu'ils soient
- Limiter les déplacements et leur logistique
- Réduire les temps d'absence liés aux déplacements
- Soutenir la montée en compétences des équipes dans tous les contextes