

Implémenter des solutions d'ingénierie des données à l'aide de Azure Databricks (DP750T00)

Durée: 4 jours

Public cible: Analyste de données

À qui s'adresse cette formation ?

Ce cours s'adresse aux ingénieurs de données ayant une connaissance fondamentale des concepts d'analytique et d'entrepôt de données, une compréhension de base du stockage informatique et des principes d'organisation des données.

Ils doivent être à l'aise avec SQL et avoir de l'expérience en Python, incluant l'utilisation de notebooks pour les tâches d'ingénierie de données.

Les apprenants doivent comprendre les espaces de travail Azure Databricks, Unity Catalog, les modèles d'accès aux données et les concepts fondamentaux d'ingénierie et d'entrepôt de données.

Une connaissance de base de la sécurité Azure, incluant Microsoft Entra ID, ainsi que des notions de contrôle de version Git est également requise.

Objectifs de la formation

À la fin de ce cours, les participants seront en mesure de :

- Provisionner et configurer un espace de travail Azure Databricks et comprendre ses composantes architecturales.
- Sélectionner et configurer les ressources de calcul appropriées selon les types de charges de travail.
- Créer et organiser des objets Unity Catalog, incluant les catalogues, schémas, tables, vues et volumes.
- Mettre en œuvre des contrôles de sécurité tels que les ACL, le filtrage de lignes, le masquage de colonnes et Azure Key Vault.
- Appliquer des pratiques de gouvernance incluant la traçabilité des données, la journalisation d'audit, les étiquettes et Delta Sharing.

- Concevoir et mettre en œuvre des modèles de données avec Delta Lake, incluant les stratégies de partitionnement et de clustering.
- Ingérer des données selon plusieurs modèles : batch, streaming, Auto Loader, COPY INTO et Lakeflow Connect.
- Nettoyer et transformer des données brutes avec SQL et PySpark, incluant les jointures, pivots et opérations de fusion.
- Mettre en œuvre des contraintes de qualité des données via l'application de schémas et les attentes Lakeflow Spark Declarative Pipeline.
- Concevoir et mettre en œuvre des pipelines multi tâches avec gestion des erreurs et logique de reprise.
- Créer et planifier des Lakeflow Jobs avec déclencheurs, alertes et redémarrages automatisés.
- Appliquer des pratiques de cycle de développement utilisant les dossiers Git et les Declarative Automation Bundles.
- Surveiller, dépanner et optimiser les charges de travail Spark à l'aide du Spark UI et d'Azure Log Analytics.

Description sommaire

DP 750T00 : Mettre en œuvre des solutions d'ingénierie de données avec Azure Databricks est un cours de quatre jours, offert en formation magistrale (ILT), qui prépare les ingénieurs de données à concevoir, bâtir et maintenir des solutions d'ingénierie de données évolutives sur Azure Databricks.- Le cours couvre l'ensemble du cycle de vie d'une charge de travail d'ingénierie de données: de la configuration de l'espace de travail et des ressources de calcul, à l'ingestion et la transformation des données, en passant par l'application de la gouvernance et le déploiement de pipelines de production.

Contenu du plan de cours

Parcours d'apprentissage 1 — Configurer un environnement Azure Databricks

-

Module 1 — Explorer Azure Databricks

-

Module 2 — Sélectionner et configurer les ressources de calcul dans Azure Databricks

-

Module 3 — Créer et organiser des objets dans Unity Catalog

Parcours d'apprentissage 2 — Sécuriser et gouverner les objets Unity Catalog dans Azure Databricks

-

Module 1 — Sécuriser les objets Unity Catalog

-

Module 2 — Gouverner les objets Unity Catalog

Parcours d'apprentissage 3 — Préparer et traiter les données avec Azure Databricks

-

Module 1 — Concevoir et mettre en œuvre la modélisation des données avec Azure Databricks

-

Module 2 — Ingérer des données dans Unity Catalog

-

Module 3 — Nettoyer, transformer et charger les données dans Unity Catalog

-

Module 4 — Mettre en œuvre et gérer les contraintes de qualité des données avec Azure Databricks

Parcours d'apprentissage 4 — Déployer et maintenir des pipelines et des charges de travail avec Azure Databricks

-

Module 1 — Concevoir et mettre en œuvre des pipelines de données avec Azure Databricks

-

Module 2 — Mettre en œuvre des Lakeflow Jobs avec Azure Databricks

-

Module 3 — Mettre en œuvre les processus de cycle de développement dans Azure Databricks

-

Module 4 — Surveiller, dépanner et optimiser les charges de travail dans Azure Databricks

Approche et méthode pédagogique

Approche pratique et structurée combinant une théorie ciblée avec des laboratoires guidés. Les participants développent progressivement des solutions d'ingénierie de données avec Azure Databricks au moyen d'exercices réels basés sur des scénarios, favorisant une application immédiate des apprentissages.

Ils apprennent à provisionner et configurer des espaces de travail Azure Databricks, à organiser et sécuriser les données avec Unity Catalog, à concevoir des modèles de données Delta Lake, à ingérer et transformer des données avec SQL et PySpark, et à mettre en œuvre des contraintes de qualité des données. Le cours met l'accent sur les pratiques modernes d'ingénierie de données, incluant les pipelines Lakeflow, les Lakeflow Jobs, les flux de développement basés sur Git et l'optimisation de performance à l'aide du Spark UI et d'Azure Log Analytics.

Animé par un formateur certifié Microsoft (MCT), le cours encourage l'interactivité et le développement de compétences techniques transférables permettant aux participants de concevoir, bâtir, sécuriser et maintenir en toute confiance des charges de travail d'ingénierie de données de niveau production sur Azure Databricks.

Prérequis

Les participants doivent posséder les connaissances et l'expérience suivantes avant de suivre ce cours :

- Connaissances fondamentales en analytique de données et concepts d'entreposage.
- Compréhension de base du stockage infonuagique et de la gestion des ressources Azure.
- Familiarité avec SQL pour interroger et gérer les données.
- Compréhension de base de Python (utilisé dans les notebooks et PySpark).
-

Compréhension des principes fondamentaux de Git et du contrôle de version.

- Connaissances de base de Microsoft Entra ID et des notions de sécurité Azure (recommandé pour le Parcours d'apprentissage 2).