Programme de formation
DP-203: Data Engineering on Microsoft Azure

(Préparation certification Microsoft DP-203)

DESCRIPTION DE LA FORMATION :

Les ingénieurs de données de Azure intègrent, transforment et consolident les données provenant de divers systèmes de données structurées et non structurées dans des structures qui conviennent à l'élaboration de solutions analytiques. Cette formation vous aidera pour y parvenir en adoptant les meilleurs pratiques.

OBJECTIFS PEDAGOGIQUES :

A l'issue de cette formation, les participants seront en capacité de :

- Concevoir une structure de stockage de données
- Concevoir une stratégie de partition
- Concevoir et mettre en œuvre des couches de service
- Mettre en œuvre des structures de stockage de données physique et de données logique
- Intégrer et transformer des données
- Concevoir et développer des solutions de traitement par lot et par flux
- Gérer les lots et les pipelines
- Concevoir la sécurité des politiques et des normes de données
- Mettre en œuvre la sécurité des données
- Surveiller le stockage et le traitement des données
- Optimiser et dépanner le stockage et le traitement des données

MÉTHODES PÉDAGOGIQUES :

- Cette formation sera principalement constituée de théorie et d'ateliers techniques qui permettront d'être rapidement opérationnel.
- Support : un support de cours officiel Microsoft en anglais sera remis aux participants au format électronique via la plateforme Skillpipe.
- Evaluation : Les acquis sont évalués tout au long de la formation par le formateur (Questions régulières, travaux pratiques, QCM ou autres méthodes).
- Formateur : le tout animé par un consultant-formateur expérimenté, nourri d'une expérience terrain, et accrédité Microsoft Certified Trainer.
- Satisfaction : à l'issue de la formation, chaque participant répond à un questionnaire d'évaluation qui est ensuite analysé en vue de maintenir et d'améliorer la qualité de nos formations.

- Suivi : une feuille d'émargement par demi-journée de présence est signée par chacun des participants.
- Cette formation peut être dispensée en format inter-entreprises ou intra-entreprise sur demande et en mode présentiel comme en distanciel.

PROGRAMME DE FORMATION :

Concevoir une structure de stockage de données

- Concevoir une solution Azure Data Lake
- Recommander des types de fichiers pour le stockage
- Recommander des types de fichiers pour les requêtes analytiques
- Conception pour une interrogation efficace et pour l'élagage des données
- Concevoir une structure de dossiers qui représente les niveaux de transformation des données
- Concevoir une stratégie de distribution concevoir une solution d'archivage de données

Concevoir une stratégie de partition

- Concevoir une stratégie de partition pour les fichiers
- Concevoir une stratégie de partition pour les charges de travail analytiques
- Concevoir une stratégie de partition pour l 'efficacité
- Concevoir une stratégie de partition pour Azure Synapse Analytics
- Identifier quand le partitionnement est nécessaire dans Azure Data Lake Storage Gen2

Concevoir la couche de service

- Concevoir des schémas en étoile
- Concevoir des dimensions qui changent lentement
- Concevoir une hiérarchie dimensionnelle
- Concevoir une solution pour les données temporelles
- Conception pour chargement incrémentiel
- Concevoir des magasins analytiques
- Concevoir des métastores dans Azure Synapse Analytics et Azure Databricks

Mettre en œuvre des structures de stockage de données physiques

- Mettre en œuvre la compression
- Implémenter le partitionnement et le sharding
- Implémenter différentes géométries de table avec les pools Azure Synapse Analytics
- Mettre en œuvre la redondance des données
- Mettre en œuvre des distributions
- Mettre en œuvre l'archivage des données

Mettre en œuvre des structures de données logiques

- Construire une solution de données temporelles
- Construire une dimension qui change lentement
- Construire une structure de dossiers logique
- Créer des tables externes
- Implémenter des structures de fichiers et de dossiers pour une interrogation et un élagage des données efficaces

Mettre en œuvre la couche de diffusion

- Fournir des données dans un schéma relationnel en étoile
- Fournir des données dans des fichiers Parquet
- Maintenir les métadonnées
- Mettre en œuvre une hiérarchie dimensionnelle

Ingérer et transformer des données

- Transformer les données à l'aide d'Apache Spark,
- Transformer les données à l'aide de Transact-SQL
- Transformer les données à l'aide de Data Factory
- Transformer les données à l'aide des pipelines Azure Synapse
- Transformer les données à l'aide de Scala
- Transformer les données à l'aide de Stream Analytics
- Nettoyer les données
- Données fractionnées
- Déchiqueter JSON
- Encoder et décoder les données
- Configurer la gestion des erreurs pour la transformation
- Normaliser et dénormaliser les valeurs
- Effectuer une analyse exploratoire des données

Concevoir et développer une solution de traitement par lots

- Développer des solutions de traitement par lots en utilisant Data Factory, Data Lake, Spark, Azure Pipelines Synapse, PolyBase et Azure Databricks
- Créer des pipelines de données
- Concevoir et mettre en œuvre des charges de données incrémentielles
- Concevoir et développer des dimensions qui changent lentement
- Gérer les exigences de sécurité et de conformité
- Mettre à l'échelle les ressources
- Configurer la taille du lot
- Concevoir et créer des tests pour les pipelines de données
- Intégrer les notebooks Jupyter / IPython dans un pipeline de données
- Gérer les données en double, manquantes ou arrivées tardivement
- Régresser à un état antérieur
- Concevoir et configurer la gestion des exceptions
- Configurer la rétention des lots
- Concevoir une solution de traitement par lots
- Déboguer les tâches Spark à l'aide de l'interface utilisateur Spark

Concevoir et développer une solution de traitement de flux

- Développer une solution de traitement de flux en utilisant Stream Analytics, Azure Databricks et Azure Event Hubs Traiter les données à l'aide du streaming structuré Spark
- Surveiller les performances et les régressions fonctionnelles
- Concevoir et créer des agrégats fenêtrés
- Gérer la dérive de schéma

- Traiter les données de séries chronologiques
- Processus à travers les partitions
- Traiter dans une partition
- Configurer les points de contrôle / le filigrane pendant le traitement
- Mettre à l'échelle les ressources
- Concevoir et créer des tests pour les pipelines de données
- Optimiser les pipelines à des fins analytiques ou transactionnelles
- Gérer les interruptions
- Concevoir et configurer la gestion des exceptions
- Relire les données de flux archivées
- Concevoir une solution de traitement de flux

Concevoir la sécurité des politiques et des normes de données

- Concevoir le cryptage pour les données au repos et en transit
- Concevoir une stratégie d'audit des données et concevoir une stratégie de masquage des données
- Concevoir une politique de conservation et de confidentialité des données
- Créer une purger des données en fonction des besoins de l'entreprise
- Concevoir le contrôle d'accès basé sur les rôles Azure (Azure RBAC) et la liste de contrôle d'accès de type POSIX (ACL) pour Data Lake Storage Gen2
- Conception de la sécurité au niveau des lignes et des colonnes

Mettre en œuvre la sécurité des données

- Masquer, crypter des données.
- Implémenter des terminaux sécurisé et la sécurité au niveau des lignes et des colonnes
- Implémenter Azure RBAC et des ACL de type POSIX pour Data Lake Storake Gen2
- Mettre en œuvre une politique de conservation et d'audit des données
- Gérer les identités, clé et secrets sur différentes plates-formes de données
- Charger un DataFrame avec des informations sensibles et gérer les informations sensibles
- Ecrire des données chiffrées dans des tables ou fichiers Parquet

Surveiller le stockage et le traitement des données

- Implémenter la journalisation utilisée par Azure Monitor
- Configurer les services de surveillance et mesurer les performances du mouvement des données
- Surveiller et mettre à jour les statistiques sur les données d'un système
- Surveiller les performances du pipeline de données et du cluster
- Mesurer les performances des requêtes
- Comprendre les options de journalisation personnalisées et planifier et surveiller les tests de pipeline
- Interpréter les métriques et les journaux Azure Monitor

Optimiser et dépanner le stockage et le traitement des données

- Réécrire les fonctions définies par l'utilisateur (UDF)
- Gérer le biais dans les données et le déversement de données
- Régler les partitions de manière aléatoire et les requêtes à l'aide d'indexeurs et du cache
- Trouver la lecture aléatoire dans un pipeline
- Optimiser la gestion des ressources

- Optimiser les pipelines à des fins analytiques ou transactionnelles et pour les charges de travail descriptives par rapport aux charges de travail analytiques
- Dépanner un travail ou une exécution ayant échoué

PRÉREQUIS :

Les candidats doivent avoir une expertise en matière d'intégration, de transformation et doivent savoir consolider divers systèmes de données structurés et non structurés dans un outils adapté à la création de solutions d'analyses. Il faut également une bonne connaissance des langages tels que SQL, Python ou Scala et comprendre l'architecture des données.

Les candidats doivent avoir suivi la formation AZ-900 Azure Fundamentals et DP-900 Data Fundamentals ou avoir un niveau d'expérience équivalent.

PRE-CERTIFCIATION :

Cette formation ouvre la porte à la certification Microsoft « DP-203 – Data Engineering on Microsoft Azure ».

DUREE : 4 jours (28 heures)

INTERLOCUTEURS : Data Engineers, Data Scientists

NIVEAU : Intermédiaire