Google Cloud Platform - Data Engineering

Référence : **GCP200DE** Durée : **4 jours** Certification : **Aucune**

CONNAISSANCES PREALABLES

• 1-Avoir suivi le cours GCP100A - Google Cloud Fundamentals: Big Data & Machine Learning ou avoir une expérience équivalente. • 2-Compétence de base avec un langage de requête commun tel que SQL. • 3-Experience avec la modélisation de données et l 'ETL. • 4-Développement d'applications à l'aide d'un langage de programmation commun tel que Python. • 5-Connaissance du machine learning et / ou des statistiques.

PROFIL DES STAGIAIRES

• Développeurs expérimentés qui sont responsables de la gestion des transformations des mégadonnées, notamment: l'extraction, le chargement, la transformation, le nettoyage et la validation des données.

OBJECTIFS

• Conception et déploiement de pipelines et d'architectures pour le traitement des données. • Création et déploiement de workflows de machine learning. • Interrogation des ensembles de données. • Visualisation des résultats des requêtes et création de rapports.

CERTIFICATION PREPAREE

Aucune

METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

FORMATEUR

Consultant-Formateur expert Cloud

METHODE D'EVALUATION DES ACQUIS

- 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

CONTENU DU COURS

Introduction à l'ingénierie des données

- Explorez le rôle d'un data engineer
- Analyser les défis d'ingénierie des données
- Introduction à BigQuery
- Data lakes et data warehouses
- Démo: requêtes fédérées avec BigQuery
- Bases de données transactionnelles vs data warehouses
- Démo: recherche de données personnelles dans votre jeu de données avec l'API DLP
- Travailler efficacement avec d'autres équipes de données
- Gérer l'accès aux données et gouvernance
- Construire des pipelines prêts pour la production
- Etude de cas d'un client GCP
- ***Lab: Analyse de données avec BigQuery***

Construire un Data Lake

- Introduction aux data lakes
- Stockage de données et options ETL sur GCP
- Construction d'un data lake à l'aide de Cloud Storage
- ***Démo: optimisation des coûts avec les classes et les fonctions cloud de Google Cloud Storage***
- Sécurisation de Cloud Storage
- Stocker tous les types de données
- ***Démo: exécution de requêtes fédérées sur des fichiers Parquet et ORC dans BigQuery***
- Cloud SQL en tant que data lake relationnel

Construire un Data Warehouse

- Le data warehouse moderne
- Introduction à BigQuery
- ***Démo: Requêter des TB + de données en quelques secondes***
- Commencer à charger des données
- ***Démo: Interroger Cloud SQL à partir de BigQuery***
- ***Lab: Chargement de données avec la console et la CLI***
- Explorer les schémas
- Exploration des jeux de données publics BigQuery avec SQL à l'aide de INFORMATION_SCHEMA
- Conception de schéma
- ***Démo: Exploration des jeux de données publics BigQuery avec SQL à l'aide de INFORMATION_SCHEMA***
- Champs imbriqués et répétés dans BigQuery
- ***Lab: tableaux et structures***
- Optimiser avec le partitionnement et le clustering
- ***Démo: Tables partitionnées et groupées dans BigQuery***
- Aperçu: Transformation de données par lots et en continu

Introduction à la construction de pipelines de données par lots EL, ELT, ETL

- Considérations de qualité
- Comment effectuer des opérations dans BigQuery
- ***Démo: ELT pour améliorer la qualité des données dans BigQuery***
- Des lacunes
- ETL pour résoudre les problèmes de qualité des données

Exécution de Spark sur Cloud Dataproc

- L'écosystème Hadoop
- Exécution de Hadoop sur Cloud Dataproc GCS au lieu de HDFS
- Optimiser Dataproc
- ***Atelier: Exécution de jobs Apache Spark sur Cloud Dataproc***

Traitement de données sans serveur avec Cloud Dataflow

- Cloud Dataflow
- Pourquoi les clients apprécient-ils Dataflow?
- Pipelines de flux de données
- ***Lab: Pipeline de flux de données simple (Python / Java)***
- ***Lab: MapReduce dans un flux de données (Python / Java)***
- ***Lab: Entrées latérales (Python / Java)***
- Templates Dataflow
- Dataflow SQL

Gestion des pipelines de données avec Cloud Data Fusion et Cloud Composer

- Création visuelle de pipelines de données par lots avec Cloud Data Fusion: composants, présentation de l'interface utilisateur, construire un pipeline, exploration de données en utilisant Wrangler
- ***Lab: Construction et exécution d'un graphe de pipeline dans Cloud Data Fusion***
- Orchestrer le travail entre les services GCP avec Cloud Composer - Apache Airflow Environment: DAG et opérateurs, planification du flux de travail
- ***Démo: Chargement de données déclenché par un événement avec Cloud Composer, Cloud Functions, Cloud Storage et BigQuery***
- ***Lab: Introduction à Cloud Composer***

Introduction au traitement de données en streaming

- Traitement des données en streaming

Serverless messaging avec Cloud Pub/Sub

- Cloud Pub/Sub
- ***Lab: Publier des données en continu dans Pub/Sub***

Fonctionnalités streaming de Cloud Dataflow

- Fonctionnalités streaming de Cloud Dataflow
- ***Lab: Pipelines de données en continu***

Fonctionnalités Streaming à haut débit BigQuery et Bigtable

- Fonctionnalités de streaming BigQuery
- ***Lab: Analyse en continu et tableaux de bord***
- Cloud Bigtable
- ***Lab: Pipelines de données en continu vers Bigtable***

Fonctionnalité avancées de BigQuery et performance

- Analytic Window Functions
- Utiliser des clauses With
- Fonctions SIG
- ***Démo: Cartographie des codes postaux à la croissance la plus rapide avec BigQuery GeoViz***
- Considérations de performance
- ***Lab: Optimisation de vos requêtes BigQuery pour la performance***
- ***Lab: Création de tables partitionnées par date dans BigQuery***

Introduction à l'analytique et à l'IA

- Qu'est-ce que l'IA ?
- De l'analyse de données ad hoc aux décisions basées sur les données
- Options pour modèles ML sur GCP

API de modèle ML prédéfinies pour les données non structurées

- Les données non structurées sont difficiles à utiliser
- API ML pour enrichir les données
- ***Lab: Utilisation de l'API en langage naturel pour classer le texte non structuré***

Big Data Analytics avec les notebooks Cloud AI Platform

- Qu'est-ce qu'un notebook
- BigQuery Magic et liens avec Pandas
- ***Lab: BigQuery dans Jupyter Labs sur IA Platform***

Pipelines de production ML avec Kubeflow

- Façons de faire du ML sur GCP

- Kubeflow AI Hub
- ***Lab: Utiliser des modèles d'IA sur Kubeflow***

Création de modèles personnalisés avec SQL dans BigQuery ML

- BigQuery ML pour la construction de modèles rapides
- ***Démo: Entraîner un modèle avec BigQuery ML pour prédire les tarifs de taxi à New York***
- Modèles pris en charge
- ***Lab: Prédire la durée d'une sortie en vélo avec un modèle de régression dans BigQuery ML***
- ***Lab: Recommandations de film dans BigQuery ML***

Création de modèles personnalisés avec Cloud AutoML

- Pourquoi Auto ML ?
- Auto ML Vision
- Auto ML NLP
- Auto ML Tables

*Notre **référent handicap** se tient à votre disposition au 01.71.19.70.30 ou par mail à referent.handicap@edugroupe.com pour recueillir vos éventuels besoins d'aménagements, afin de vous offrir la meilleure expérience possible.*