

Introduction à l'ingénierie des données sur Google Cloud

Référence : **GCP100DE**

Durée : **1 jour (7 heures)**

Certification : **Aucune**

CONNAISSANCES PRÉALABLE

- Compréhension des principes d'ingénierie des données, y compris les processus ETL/ELT, la modélisation des données et les formats de données courants (Avro, Parquet, JSON)
- Familiarité avec les concepts d'architecture de données, en particulier les entrepôts de données (Data Warehouses) et les lacs de données (Data Lakes)
- Maîtrise de SQL pour l'interrogation des données
- Maîtrise d'un langage de programmation courant (Python recommandé)
- Familiarité avec l'utilisation des interfaces de ligne de commande (CLI)
- Familiarité avec les concepts et services de base de Google Cloud (Compute, Storage et gestion des identités)

PROFIL DES STAGIAIRES

- Ingénieurs de données, Analystes de données, Architectes de données

OBJECTIFS

- Concevoir des systèmes de traitement de données évolutifs dans Google Cloud
- Construire et gérer des pipelines de données robustes en streaming et en batch
- Différencier les architectures de données et implémenter les concepts de lakehouse et de pipelines de données
- Utiliser les outils IA/ML pour optimiser les performances et obtenir des informations sur les processus et les données

CERTIFICATION PRÉPARÉE

- Aucune

MÉTHODES PÉDAGOGIQUES

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

FORMATEUR

- Consultant-formateur expert de GCP

MÉTHODES D'ÉVALUATION 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

ACCESSIBILITÉ DE LA FORMATION

- EduGroupe met en place un ensemble de dispositifs pour accueillir, accompagner et adapter ses formations aux personnes en situation de handicap (PSH).
- Notre référent(e) handicap se tient à votre disposition au 01.71.19.70.30 ou par mail à referent.handicap@edugroupe.com pour tout besoin d'aménagement, afin de vous offrir la meilleure expérience possible.

CONTENU DU COURS

1. Tâches et composants de l'ingénierie des données

- Le rôle d'un ingénieur de données
- Formats de données
- Options de gestion des métadonnées sur Google Cloud
- Sources de données versus récepteurs de données
- Options de solutions de stockage sur Google Cloud
- Partage de jeux de données avec Analytics Hub
- 💡 *Lab : Chargement de données dans BigQuery*
- 💡 *Quiz*

2. Réplication et migration des données

- Architecture de réplication et de migration
- Déplacement de jeux de données
- L'outil de ligne de commande gcloud
- Datastream

3. Le modèle de pipeline de données d'extraction et de changement

- Architecture d'extraction et de chargement
- Service de transfert de données BigQuery
- L'outil de ligne de commande bq
- BigLake
- 💡 *Lab : BigLake : Démarrage rapide*
- 💡 *Quiz*

4. Le modèle de pipeline de données d'extraction, de chargement et de transformation

- Architecture d'extraction, de chargement et de transformation (ELT)
- Dataform
- Scripting SQL et planification avec BigQuery
- 💡 *Lab : Créer et exécuter un workflow SQL dans Dataform*
- 💡 *Quiz*

5. Le modèle de pipeline de données d'extraction, de transformation et de chargement

- Architecture d'extraction, de transformation et de chargement (ETL)
- Traitement de données en batch avec Dataproc
- Bigtable et pipelines de données
- Outils d'interface graphique Google Cloud pour les pipelines de données ETL
- Options de traitement de données en streaming
- 💡 *Lab : Utiliser Dataproc Serverless pour Spark pour charger BigQuery (optionnel)*
- 💡 *Lab : Créer un pipeline de données en streaming pour un tableau de bord en temps réel avec Dataflow*
- 💡 *Quiz*

6. Techniques d'automatisation

- Modèles d'automatisation et options pour les pipelines
- Cloud Composer
- Eventarc
- Cloud Scheduler et Workflows
- Cloud Run Functions
- 💡 *Lab : Utiliser Cloud Run Functions pour charger BigQuery (optionnel)*
- 💡 *Quiz*