

# Data Engineering on Google Cloud Platform

Les compétences essentielles pour concevoir et construire des systèmes de traitement de données et des workflows de machine learning sur Google Cloud

4 jour(s) / 28h

## Présentation du cours

Ce cours de quatre jours dirigé par un instructeur offre aux participants une introduction pratique à la conception et à la création de systèmes de traitement des données sur Google Cloud Platform. Grâce à une combinaison de présentations, de démonstrations et de travaux pratiques, les participants apprendront à concevoir des systèmes de traitement des données, à construire des pipelines de données de bout en bout, à analyser les données et à effectuer un apprentissage automatique. Le cours couvre les données structurées, non structurées et en streaming.

## Objectifs pédagogiques

- 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

## Public cible

- 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

## Prérequis

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

## Programme

### Module 1: 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

### Module 2: 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

### **Module 3: Construire un Data Warehouse**

- Le data warehouse moderne
- Introduction à BigQuery
- Démo: Requête 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

### **Module 4: 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

### **Module 5: 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

## **Module 6: 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

## **Module 7: 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

## **Module 8: Introduction au traitement de données en streaming**

- Traitement des données en streaming

## **Module 9: Serverless messaging avec Cloud Pub/Sub**

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

## **Module 10: Fonctionnalités streaming de Cloud Dataflow**

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

## **Module 11: 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

## **Module 12: 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

### **Module 13: 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

### **Module 14: 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

### **Module 15: 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

### **Module 16: Pipelines de production ML avec Kubeflow**

- Façons de faire du ML sur GCP
- Kubeflow AI Hub
- Lab: Utiliser des modèles d'IA sur Kubeflow

### **Module 17: 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

### **Module 18: Création de modèles personnalisés avec Cloud AutoML**

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