

Google Cloud Platform – Data Engineering

4 jours (28 heures)

Perfectionnement

Classe à distance

à partir de

3 160 € HT par personne (TVA en vigueur)

Contenu mis à jour le 25/09/2026. Document téléchargé le 02/10/2026.

Objectifs

À l'issue de cette formation, vous serez capable de :

- Concevoir et créer des systèmes de traitement de données sur Google Cloud
- Traiter des données en batch et en streaming en implémentant des pipelines de données à l'autoscaling sur Dataflow
- Tirer des informations commerciales à partir de très grands ensembles de données à l'aide de BigQuery
- Exploiter les données non structurées à l'aide de Spark et des API ML sur Dataproc
- Mettre en oeuvre des analyses instantanées à partir des données en streaming.

Les compétences visées par cette formation font l'objet d'une évaluation.

À qui s'adresse cette formation ?

Ingénieurs de données, administrateurs de bases de données et de systèmes.

Prérequis

Avoir une expérience préalable de Google Cloud en utilisant Cloud Shell et en accédant aux produits depuis la console Google Cloud. Avoir une maîtrise de base d'un langage de requête tel que le SQL. Avoir une expérience en modélisation de données et en activités ETL (extraction, transformation, chargement), ainsi qu'une expérience dans le développement d'application à l'aide d'un langage de programmation courant tel que Python.

Programme

1 Pendant votre formation

Google Cloud Platform – Data Engineering

Tâches et composants de l'ingénierie des données

- Expliquer le rôle d'un ingénieur de données
- Comprendre les différences entre une source de données et un récepteur de données
- Expliquer les différents types de formats de données
- Expliquer les options de solutions de stockage sur Google Cloud
- En savoir plus sur les options de gestion des métadonnées sur Google Cloud
- Comprendre comment partager facilement des ensembles de données avec Analytics Hub
- Comprendre comment charger des données dans BigQuery à l'aide de la console Google Cloud et/ou de la CLI gcloud

Lab

- *Chargement des données dans BigQuery*

Réplication et migration de données

- Expliquer l'architecture de base de réplication et de migration de données dans Google Cloud
- Comprendre les options et les cas d'utilisation de l'outil de ligne de commande gcloud
- Expliquer la fonctionnalité et les cas d'utilisation du service de transfert de stockage
- Expliquer la fonctionnalité et les cas d'utilisation de l'appliance de transfert
- Comprendre les fonctionnalités et le déploiement de Datastream

Lab

- *Datastream : réplication de PostgreSQL vers BigQuery*

Le modèle de pipeline de données d'extraction et de chargement

- Expliquer le schéma d'architecture de base d'extraction et de chargement
- Comprendre les options de l'outil en ligne de commande bq
- Expliquer la fonctionnalité et les cas d'utilisation du service de transfert de données BigQuery
- Expliquer la fonctionnalité et les cas d'utilisation de BigLake en tant que modèle sans extraction-chargement

Lab

- *BigLake : démarrage rapide*

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

- Expliquer le schéma d'architecture de base d'extraction, de chargement et de transformation
- Comprendre un pipeline ELT commun sur Google Cloud

- En savoir plus sur les capacités de scripting SQL et de planification de BigQuery
- Expliquer la fonctionnalité et les cas d'utilisation de Dataform

Lab

- *Créer et exécuter un workflow SQL dans Dataform*

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

- Expliquer le schéma d'architecture de base d'extraction, de transformation et de chargement
- En savoir plus sur les outils d'interface graphique sur Google Cloud utilisés pour les pipelines de données ETL
- Expliquer le traitement des données en batch avec Dataproc
- Apprendre à utiliser Dataproc sans serveur pour Spark pour l'ETL
- Expliquer les options de traitement des données en streaming
- Expliquer le rôle que joue BigTable dans les pipelines de données

Labs

- *Utiliser Dataproc sans serveur pour Spark pour charger BigQuery*
- *Créer une pipeline de données en streaming pour un tableau de bord en temps réel avec Dataflow*

Techniques d'automatisation

- Expliquer les modèles d'automatisation et les options disponibles pour les pipelines
- En savoir plus sur Google Scheduler et les workflows
- En savoir plus sur Cloud Composer
- En savoir plus sur les fonctions Cloud Run
- Expliquer la fonctionnalité et les cas d'utilisation d'automatisation pour Eventarc

Lab

- *Utiliser les fonctions Cloud Run pour charger BigQuery*

Introduction à l'ingénierie des données

- Discuter des défis de l'ingénierie des données et de la manière dont la création de pipelines de données dans le Cloud aide à les résoudre
- Examiner et comprendre le but d'un lac de données par rapport à un entrepôt de données, et quand utiliser lequel

Lab

- *Utiliser BigQuery pour faire des analyses*

Construire un lac de données

- Discuter pourquoi Cloud Storage est une excellente option pour construire un lac de données sur Google Cloud
- Expliquer comment utiliser Cloud SQL pour un lac de données relationnel

Lab

- *Chargement des données de taxi dans Cloud SQL*

Construire un entrepôt de données

- Discuter des exigences d'un entrepôt moderne

- Expliquer pourquoi BigQuery est la solution d'entreposage de données évolutive sur Google Cloud
- Discuter des concepts de base de BigQuery et examiner les options de chargement des données dans BigQuery

Labs

- *Travailler avec des données JSON et des tableaux dans BigQuery*
- *Tables partitionnées dans BigQuery*

Introduction à la création de pipelines de données en batch

- Examiner différentes méthodes de chargement de données dans vos lacs et entrepôts de données :
 - EL
 - ELT
 - ETL

Exécuter Spark sur Dataproc

- Examiner l'écosystème Hadoop
- Discuter de la manière de migrer (lift and shift) vos charges de travail Hadoop existantes vers le Cloud à l'aide de Dataproc
- Expliquer quand utiliser Cloud Storage au lieu du stockage HDFS
- Expliquer comment optimiser les tâches Dataproc

Lab

- *Exécuter des tâches Apache Spark sur Dataproc*

Traitement de données sans serveur avec Dataflow

- Identifier les fonctionnalités que les clients apprécient dans Dataflow
- Discuter des concepts de base de Dataflow
- Examiner l'utilisation des modèles Dataflow et SQL
- Ecrire un pipeline Dataflow simple et l'exécuter à la fois localement et sur le Cloud
- Identifier les opérations Map et Reduce, exécuter le pipeline et utiliser les paramètres de ligne de commande
- Lire les données de BigQuery dans Dataflow et utiliser la sortie d'un pipeline comme entrée secondaire pour un autre pipeline

Labs

- *Un pipeline Dataflow simple (Python / Java)*
- *MapReduce dans Beam (Python / Java)*
- *Entrées secondaires (Python / Java)*

Gérer les pipelines de données avec Cloud Data Fusion et Cloud Composer

- Discuter de la manière de gérer vos pipelines de données avec Cloud Data Fusion et Cloud Composer
- Résumer comment Cloud Data Fusion permet aux analystes de données et aux développeurs ETL de manipuler les données et de créer des pipelines de manière visuelle
- Décrire comment Cloud Composer peut aider à orchestrer le travail sur plusieurs services Google Cloud

Labs

- *Créer et exécuter un graphe de pipeline dans Data Fusion*
- *Une introduction à Cloud Composer*

Introduction au traitement des données en streaming

- Expliquer le traitement des données en streaming
- Identifier les produits et outils Google Cloud qui peuvent aider à relever les défis des données en streaming

Messagerie sans serveur avec Pub/Sub

- Décrire le service Pub/Sub
- Expliquer le fonctionnement de Pub/Sub
- Simuler des données de capteurs en streaming en temps réel à l'aide de Pub/Sub

Lab

- *Publier des données en streaming dans Pub/Sub*

Fonctionnalités de streaming de Dataflow

- Décrire le service Dataflow
- Construire un pipeline de traitement de flux pour les données de trafic en direct
- Démontrer comment gérer les données en retard à l'aide de filigranes, de déclencheurs et d'accumulation

Lab

- *Pipelines de données en streaming*

Fonctionnalités de streaming à haut débit de BigQuery et BigTable

- Décrire comment effectuer une analyse ad hoc sur les données en streaming à l'aide de BigQuery et des tableaux de bord
- Discuter de BigTable en tant que solution à faible latence
- Décrire comment architecturer pour BigTable et comment ingérer des données dans BigTable
- Mettre en évidence les considérations de performance pour les services concernés

Labs

- *Analyses en streaming et tableaux de bord*
- *Générer du contenu d'e-mail personnalisé avec les requêtes continues de BigQuery et Gemini*
- *Pipelines de données en streaming dans BigTable*

Fonctionnalités avancées et performances de BigQuery

- Passer en revue certaines des capacités d'analyse avancées de BigQuery
- Discuter des moyens d'améliorer les performances des requêtes

Lab

- *Optimiser vos requêtes BigQuery pour la performance*

Financement

Plusieurs dispositifs de financement sont accessibles via les OPCO (Opérateurs de Compétences), organismes agréés par le ministère du Travail dont le rôle est d'accompagner, collecter et gérer les contributions des entreprises au titre du financement de la formation professionnelle.

Pour plus d'information, une équipe de gestionnaires SKOLAE Formation spécialisée vous accompagne dans le choix de vos formations et la gestion administrative.

[Nous contacter pour en savoir plus →](#)

Modalités pédagogiques

Formation délivrée en présentiel ou distanciel* (blended-learning, e-learning, classe virtuelle, présentiel à distance).

Le formateur alterne entre méthode** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation).

Variables suivant les formations, les moyens pédagogiques mis en oeuvre sont :

- Ordinateurs Mac ou PC (sauf pour certains cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel)
- Environnements de formation installés sur les postes de travail ou en ligne
- Supports de cours et exercices

En cas de formation intra sur site externe à M2I, le client s'assure et s'engage également à avoir toutes les ressources matérielles pédagogiques nécessaires (équipements informatiques...) au bon déroulement de l'action de formation visée conformément aux prérequis indiqués dans le programme de formation communiqué.

* nous consulter pour la faisabilité en distanciel

** ratio variable selon le cours suivi

[Accessibilité aux personnes en situation de handicap](#)

Sessions

À distance

Du 26 octobre 2026 au 29 octobre 2026