Ingenierie des donnees sur Google Cloud

Prix : 3 160 € HT

Durée : 4 jours

Code de Référence : GCP200DE

Catalogue Google Cloud Platform

Cette formation cours utilise des presentations, des demonstrations et des travaux pratiques pour vous montrer comment concevoir des systemes de traitement de donnees, creer des pipelines de donnees de bout en bout, analyser des donnees et implementer le machine learning.

Objectifs de la formation

Cette formation Google Cloud permet aux participants d'acquérir les compétences suivantes:

- Concevoir des systemes de traitement de donnees evolutifs dans Google Cloud
- Differencier les architectures de donnees et implementer les concepts de lakehouse et de pipelines de donnees
- Construire et gerer des pipelines de donnees robustes en streaming et en batch
- Utiliser les outils IA/ML pour optimiser les performances et obtenir des informations sur les processus et les donnees

Public

Cette formation Google Cloud s'adresse aux ingenieurs de donnees, analystes de donnees, architectes de données.

Cette formation est accessible aux personnes en situation de handicap, contactez-nous pour en savoir plus.

Prérequis

Pour tirer le meilleur parti de ce cours, les participants doivent disposer des éléments suivants :

- Comprehension des principes d'ingenierie des donnees, y compris les processus ETL/ELT, la modelisation des donnees et les formats de donnees courants (Avro, Parquet, JSON)
- Familiarite avec les concepts d'architecture de donnees, en particulier les entrepots de donnees (Data Warehouses) et les lacs de donnees (Data Lakes)
- Maitrise de SQL pour l'interrogation des donnees
- Maitrise d'un langage de programmation courant (Python recommande)
- Familiarite avec l'utilisation des interfaces de ligne de commande (CLI)

- Familiarite avec les concepts et services de base de Google Cloud (Compute, Storage et gestion des identites)

Vous souhaitez faire vérifier vos prérequis ? Contactez-nous pour l'organisation d'un entretien téléphonique avec un de nos consultants formateurs.

Pour une efficacité renforcée, le nombre de participants est limitée à 12. Le maintien des sessions est conditionné à un minimum de 3 participants.

Programme de la formation

Module 1 : Taches et composants de l'ingénierie des données

- Expliquer le role d'un ingenieur de donnees
- Comprendre les differences entre une source de donnees et un recepteur de donnees
- Expliquer les differents types de formats de donnees
- Expliquer les options de solutions de stockage sur Google Cloud
- Decouvrir les options de gestion des metadonnees sur Google Cloud
- Comprendre comment partager facilement des jeux de donnees avec Analytics Hub
- Comprendre comment charger des donnees dans BigQuery a l'aide de la console Google Cloud ou de la CLI gcloud

Lab : Chargement de donnees dans BigQuery
Quiz

Module 2 : Réplication et migration de données

- Expliquer l'architecture de base de replication et de migration de donnees de Google Cloud
- Comprendre les options et les cas d'utilisation de l'outil de ligne de commande gcloud
- Expliquer la fonctionnalite et les cas d'utilisation de Storage Transfer Service
- Expliquer la fonctionnalite et les cas d'utilisation de Transfer Appliance
- Comprendre les fonctionnalites et le deploiement de Datastream

Module 3 : Le modèle de pipeline de données d'extraction et de chargement

- Expliquer le schema d'architecture de base d'extraction et de chargement
- Comprendre les options de l'outil de ligne de commande bq
- Expliquer la fonctionnalite et les cas d'utilisation du service de transfert de donnees BigQuery
- Expliquer la fonctionnalite et les cas d'utilisation de BigLake en tant que modele sans extraction-chargement

Lab : BigLake : Demarrage rapide
Quiz

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

- Expliquer le schema d'architecture de base d'extraction, de chargement et de transformation
- Comprendre un pipeline ELT courant sur Google Cloud
- Decouvrir les capacites de scripting SQL et de planification de BigQuery
- Expliquer la fonctionnalite et les cas d'utilisation de Dataform

Lab : Creer et executer un workflow SQL dans Dataform
Quiz

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

- Expliquer le schema d'architecture de base d'extraction, de transformation et de chargement
- Decouvrir les outils d'interface graphique sur Google Cloud utilises pour les pipelines de donnees ETL
- Expliquer le traitement des donnees en batch avec Dataproc
- Apprendre a utiliser Dataproc Serverless pour Spark pour l'ETL
- Expliquer les options de traitement des donnees en streaming
- Expliquer le role que joue Bigtable dans les pipelines de donnees

Lab : Utiliser Dataproc Serverless pour Spark pour charger BigQuery (optionnel)
Lab : Creer un pipeline de donnees en streaming pour un tableau de bord en temps reel avec Dataflow
Quiz

Module 6 : Techniques d'automatisation

- Expliquer les modeles d'automatisation et les options disponibles pour les pipelines
- Decouvrir Cloud Scheduler et Workflows
- Decouvrir Cloud Composer
- Decouvrir Cloud Run Functions
- Expliquer la fonctionnalite et les cas d'utilisation d'automatisation pour Eventarc

Lab : Utiliser Cloud Run Functions pour charger BigQuery (optionnel)
Quiz

Module 7 : Introduction à l'ingénierie des données moderne sur Google Cloud

- Comparer et contraster les architectures de lac de donnees, d'entrepot de donnees et de lakehouse de donnees
- Evaluer les avantages de l'approche lakehouse

Quiz

Module 8 : Construire un lakehouse de données avec Cloud Storage, les formats ouverts et BigQuery

- Discuter des options de stockage de donnees, y compris Cloud Storage pour les fichiers, les formats de table ouverts comme Apache Iceberg, BigQuery pour les donnees analytiques et AlloyDB pour les donnees operationnelles
- Comprendre le role d'AlloyDB pour les cas d'utilisation de donnees operationnelles

Quiz
Lab : Requete federee avec BigQuery

Module 9 : Moderniser les entrepots de données avec BigQuery et BigLake

- Expliquer pourquoi BigQuery est une solution d'entreposage de donnees evolutive sur Google Cloud

- Discuter des concepts de base de BigQuery
- Comprendre le role de BigLake dans la creation d'une architecture lakehouse unifiee et son integration avec BigQuery pour les donnees externes
- Apprendre comment BigQuery interagit nativement avec les tables Apache Iceberg via BigLake

Quiz
Lab : Interroger des donnees externes et des tables Iceberg

Module 10 : Modèles avancés de lakehouse et gouvernance des données

- Implementer des pratiques robustes de gouvernance et de securite des donnees sur la plateforme de donnees unifiee, y compris la protection des donnees sensibles et la gestion des metadonnees
- Explorer l'analytique avancee et le machine learning directement sur les donnees du lakehouse

Quiz

Module 11 : Labs et bonnes pratiques

- Renforcer les principes fondamentaux de la plateforme de donnees de Google Cloud

Lab : Demarrer avec BigQuery ML
Lab : Recherche vectorielle avec BigQuery

Module 12 : Quand choisir les pipelines de données en batch

- Expliquer le role critique d'un ingenieur de donnees dans le developpement et la maintenance des pipelines de donnees en batch
- Decrire les composants de base et le cycle de vie typique des pipelines de donnees en batch, de l'ingestion a la consommation en aval
- Analyser les defis courants du traitement de donnees en batch, tels que le volume de donnees, la qualite, la complexite et la fiabilite, et identifier les services Google Cloud cles qui peuvent les resoudre

Quiz

Module 13 : Concevoir et construire des pipelines de données en batch évolutifs

- Concevoir des pipelines de donnees en batch evolutifs pour l'ingestion et la transformation de donnees a haut volume
- Optimiser les jobs en batch pour un haut debit et une efficacite des couts en utilisant diverses techniques de gestion des ressources et d'ajustement des performances

Quiz
Lab : Construire un pipeline de donnees en batch simple avec Serverless pour Apache Spark (optionnel)
Lab : Construire un pipeline de donnees en batch simple avec l'interface Dataflow Job Builder (optionnel)

Module 14 : Controler la qualité des données dans les pipelines de données en batch

- Developper des regles de validation des donnees et une logique de nettoyage pour assurer la qualite des donnees dans les pipelines en batch
- Implementer des strategies pour gerer l'evolution des schemas et effectuer la deduplication des donnees dans les grands jeux de donnees

Lab : Valider la qualite des donnees dans un pipeline en batch avec Serverless pour Apache Spark (optionnel)
Quiz

Module 15 : Orchestrer et surveiller les pipelines de données en batch

- Orchestrer des workflows de pipelines de donnees en batch complexes pour une planification efficace et un suivi de lignage
- Implementer une gestion robuste des erreurs, une surveillance et une observabilite pour les pipelines de donnees en batch

Lab : Construire des pipelines en batch dans Cloud Data Fusion
Quiz

Module 16 : Construire des pipelines de donnees en streaming sur Google Cloud

- Introduire les objectifs d'apprentissage du cours et le scenario qui sera utilise pour apporter un apprentissage pratique a la construction de pipelines de donnees en streaming
- Decrire le concept de pipelines de donnees en streaming, les defis associes et le role de ces pipelines dans le processus d'ingenierie des donnees

Lab : Utiliser des modèles d'IA sur Kubeflow

Module 17 : Cas d'utilisation du streaming et architectures de référence

- Comprendre les differents cas d'utilisation du streaming et leurs applications, y compris le Streaming ETL, le Streaming IA/ML, les applications de streaming et le Reverse ETL
- Identifier et decrire les architectures types courantes pour les donnees en streaming, y compris le Streaming ETL, le Streaming IA/ML, les applications de streaming et le Reverse ETL

Quiz

Module 18 : Plongée appronfondie dans les produits

- Pub/Sub et Managed Service for Apache Kafka : Definir les concepts de messagerie, savoir quand utiliser Pub/Sub ou Managed Service for Apache Kafka
- Dataflow : Decrire le service et les defis avec les donnees en streaming, construire et deployer un pipeline de streaming
- BigQuery : Explorer les differentes methodes d'ingestion de donnees, utiliser les requetes continues BigQuery, BigQuery ETL et le reverse ETL, configurer le streaming Pub/Sub vers BigQuery, architecturer les pipelines de streaming BigQuery
- Bigtable : Decrire la vue d'ensemble du mouvement et de l'interaction des donnees, etablir un pipeline de streaming de Dataflow vers Bigtable, analyser le flux de donnees continu Bigtable pour les tendances avec BigQuery, synchroniser l'analyse des tendances dans l'application utilisateur

Lab : Streamer des donnees avec des pipelines – Cas d’utilisation Esports (optionnel)
Quiz
Lab : Utiliser Apache Beam et Bigtable pour enrichir les donnees de contenu telechargeables (DLC) esports
Quiz
Lab : Streamer des donnees e-sports avec Pub/Sub et BigQuery
Quiz
Lab : Surveiller le chat e-sports avec Streamlit
Quiz

Méthodes pédagogiques

Des exercices pratiques et des démonstrations vous permettront de mettre en pratique les notions théoriques présentées.

Méthodes d’évaluation des acquis

Afin d’évaluer l’acquisition de vos connaissances et compétences, il vous sera envoyé un formulaire d’auto-évaluation, qui sera à compléter en amont et à l’issue de la formation.
Un certificat de réalisation de fin de formation est remis au stagiaire lui permettant de faire valoir le suivi de la formation.