Ingénierie des données sur Google Cloud

Durée: 4 Jours **Réf de cours: GO5975** **Méthodes d'apprentissage: Intra-entreprise & sur-mesure**

Résumé:

Acquérir une expérience pratique dans la conception et la création de systèmes de traitement de données sur Google Cloud.
Ce cours utilise des présentations, des démonstrations et des travaux pratiques pour vous montrer comment concevoir des systèmes de traitement de données, créer des pipelines de données de bout en bout, analyser des données et implémenter le machine learning dans Google Cloud.
Ce cours couvre les données structurées, non structurées et en streaming.
Ce cours est composé des quatre modules suivants : Introduction à l'ingénierie des données sur Google CloudConstruire des lacs de données et des entrepôts de données avec Google CloudConstruire des pipelines de données en batch sur Google CloudConstruire des pipelines de données en streaming sur Google Cloud
Mis à jour 13/03/2026

Formation intra-entreprise

Cette formation est délivrable en session intra-entreprise, dans vos locaux ou dans les nôtres. Son contenu peut être adapté sur-mesure pour répondre aux besoins de vos collaborateurs. Contactez votre conseiller formation Global Knowledge ou adressez votre demande à info@globalknowledge.fr.

Public visé:

Ingénieurs de données, Analystes de données, Architectes de données.

Objectifs pédagogiques:

- A l'issue de la formation, les participants seront capables de :
- Concevoir des systèmes de traitement de données évolutifs dans Google Cloud.
- Différencier les architectures de données et implémenter les concepts de lakehouse et de pipelines de données.
- Construire et gérer des pipelines de données robustes en streaming et en batch.
- Utiliser les outils IA/ML pour optimiser les performances et obtenir des informations sur les processus et les données.

Pré-requis:

- 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).
- Maîtrise de SQL pour l'interrogation des données.
- Maîtrise d'un langage de programmation courant (Python recommandé).
- Familiarité avec :
- les concepts d'architecture de données, en particulier les entrepôts de données (Data Warehouses) et les lacs de données (Data Lakes).
- l'utilisation des interfaces de ligne de commande (CLI).
- les concepts et services de base de Google Cloud (Compute, Storage et gestion des identités).

Test et certification

Après cette formation, nous vous conseillons le(s) module(s) suivant(s):

N/A

Contenu:

Séquence 1 : Introduction a l'ingénierie des données sur Google Cloud

Module 1 :

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
- Découvrir les options de gestion des métadonnées sur Google Cloud
- Comprendre comment partager facilement des jeux de données avec Analytics Hub
- Comprendre comment charger des données dans BigQuery à l'aide de la console Google Cloud ou de la CLI gcloud
- Exercice pratique : Chargement de donnees dans BigQuery

Module 2: Réplication et migration des données

- Expliquer l'architecture de base de réplication et de migration de données de 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 de Storage Transfer Service
- Expliquer la fonctionnalité et les cas d'utilisation de Transfer Appliance
- Comprendre les fonctionnalités et le déploiement de Datastream

Module 3 : 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 de 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
- Exercice pratique BigLake : Démarrage rapide

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

Séquence 2 : Construire des lacs de données et des entrepôts de données avec Google Cloud

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

- Comparer et contraster les architectures de lac de données, d'entrepôt de données et de lakehouse de données.
- Évaluer les avantages de l'approche lakehouse.

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

- Discuter des options de stockage de données, y compris Cloud Storage pour les fichiers, les formats de table ouverts comme Apache Iceberg, BigQuery pour les données analytiques et AlloyDB pour les données opérationnelles.
- Comprendre le rôle d'AlloyDB pour les cas d'utilisation de données opérationnelles.
- Exercice pratique : Requête fédérée avec BigQuery

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

- Expliquer pourquoi BigQuery est une solution d'entreposage de données évolutive sur Google Cloud.
- Discuter des concepts de base de BigQuery.
- Comprendre le rôle de BigLake dans la création d'une architecture lakehouse unifiée et son intégration avec BigQuery pour les données externes.
- Apprendre comment BigQuery interagit nativement avec les tables Apache Iceberg via BigLake.
- Exercice pratique : Interroger des données externes et des tables Iceberg

Module 10 : Modèles avancés de lakehouse et de gouvernance des données

- Implémenter des pratiques robustes de gouvernance et de sécurité des données sur la plateforme de données unifiée, y compris la protection des données sensibles et la gestion des métadonnées.
- Explorer l'analytique avancée et le machine learning directement sur les données du lakehouse.

Module 11 : Exercices pratiques et bonnes pratiques

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

- Concevoir des pipelines de données en batch évolutifs pour l'ingestion et la transformation de données à haut volume.
- Optimiser les jobs en batch pour un haut débit et une efficacité des coûts en utilisant diverses techniques de gestion des ressources et d'ajustement des performances.
- Exercice pratique : Construire un pipeline de données en batch simple avec Serverless pour Apache Spark
- Exercice pratique : Construire un pipeline de données en batch simple avec l'interface Dataflow Job Builder

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

- Développer des règles de validation des données et une logique de nettoyage pour assurer la qualité des données dans les pipelines en batch.
- Implémenter des stratégies pour gérer l'évolution des schémas et effectuer la déduplication des données dans les grands jeux de données.
- Exercice pratique : Valider la qualité des données dans un pipeline en batch avec Serverless pour Apache Spark

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

- Orchestrer des workflows de pipelines de données en batch complexes pour une planification efficace et un suivi de lignage.
- Implémenter une gestion robuste des erreurs, une surveillance et une observabilité pour les pipelines de données en batch.
- Exercice pratique : Construire des pipelines en batch dans Cloud Data Fusion

Séquence 4 : Construire des pipelines de données en streaming sur Google Cloud

Module 16 : Concepts de pipeline de données

- Introduire les objectifs d'apprentissage du cours et le scénario qui sera utilisé pour apporter un apprentissage pratique à la construction de pipelines de données en streaming.
- Décrire le concept de pipelines de données en streaming, les défis associés et le rôle de ces pipelines dans le processus d'ingénierie des données.

Module 17 : Cas d'utilisation du streaming et

- Expliquer le schéma d’architecture de base d’extraction, de chargement et de transformation
- Comprendre un pipeline ELT courant sur Google Cloud
- Découvrir les capacités de scripting SQL et de planification de BigQuery
- Expliquer la fonctionnalité et les cas d’utilisation de Dataform
- Exercice pratique : Créer et exécuter un workflow SQL dans Dataform

Module 5 : 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
- Découvrir 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 Serverless 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
- Exercice pratique : Utiliser Dataproc Serverless pour Spark pour charger BigQuery
- Exercice pratique : Créer un pipeline de données en streaming pour un tableau de bord en temps réel avec Dataflow

Module 6 : Techniques d'automatisation

- Expliquer les modèles d’automatisation et les options disponibles pour les pipelines
- Découvrir Cloud Scheduler et Workflows
- Découvrir Cloud Composer
- Découvrir Cloud Run Functions
- Expliquer la fonctionnalité et les cas d’utilisation d’automatisation pour Eventarc
- Exercice pratique : Utiliser Cloud Run Functions pour charger BigQuery
- Renforcer les principes fondamentaux de la plateforme de données de Google Cloud.
- Exercice pratique : Démarrer avec BigQuery ML
- Exercice pratique : Recherche vectorielle avec BigQuery

Séquence 3 : Construire des pipelines de données en batch sur Google Cloud

Module 12 : Déterminer quand choisir les pipelines de données en batch

- Expliquer le rôle critique d'un ingénieur de données dans le développement et la maintenance des pipelines de données en batch.
- Décrire les composants de base et le cycle de vie typique des pipelines de données en batch, de l'ingestion à la consommation en aval.
- Analyser les défis courants du traitement de données en batch, tels que le volume de données, la qualité, la complexité et la fiabilité, et identifier les services Google Cloud clés qui peuvent les résoudre.

architectures de référence

- Comprendre les différents 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 décrire les architectures types courantes pour les données en streaming, y compris le Streaming ETL, le Streaming IA/ML, les applications de streaming et le Reverse ETL.

Module 18 : Plongée approfondie (Deep Dive) dans les produits

- Pub/Sub et Managed Service for Apache Kafka : Définir les concepts de messagerie, savoir quand utiliser Pub/Sub ou Managed Service for Apache Kafka.
- Dataflow : Décrire le service et les défis avec les données en streaming, construire et déployer un pipeline de streaming.
- BigQuery : Explorer les différentes méthodes d'ingestion de données, utiliser les requêtes continues BigQuery, BigQuery ETL et le reverse ETL, configurer le streaming Pub/Sub vers BigQuery, architecturer les pipelines de streaming BigQuery.
- Bigtable : Décrire la vue d'ensemble du mouvement et de l'interaction des données, établir un pipeline de streaming de Dataflow vers Bigtable, analyser le flux de données continu Bigtable pour les tendances avec BigQuery, synchroniser l'analyse des tendances dans l'application utilisateur.
- Exercice pratique : Streamer des données avec des pipelines - Cas d'utilisation Esports
- Exercice pratique : Utiliser Apache Beam et Bigtable pour enrichir les données de contenu téléchargeables (DLC) esports
- Exercice pratique : Streamer des données e-sports avec Pub/Sub et BigQuery

Module 19 : Revue des apprentissages du cours

Méthodes pédagogiques :

Un support de cours officiel sera fourni aux participants.

Autres moyens pédagogiques et de suivi:

• Compétence du formateur : Les experts qui animent la formation sont des spécialistes des matières abordées et ont au minimum cinq ans d'expérience d'animation. Nos équipes ont validé à la fois leurs connaissances techniques (certifications le cas échéant) ainsi que leur compétence pédagogique.
• Suivi d'exécution : Une feuille d'émargement par demi-journée de présence est signée par tous les participants et le formateur.
• En fin de formation, le participant est invité à s'auto-évaluer sur l'atteinte des objectifs énoncés, et à répondre à un questionnaire de satisfaction qui sera ensuite étudié par nos équipes pédagogiques en vue de maintenir et d'améliorer la qualité de nos prestations.

Délais d'inscription :

• Vous pouvez vous inscrire sur l'une de nos sessions planifiées en inter-entreprises jusqu'à 5 jours ouvrés avant le début de la formation sous réserve de disponibilité de places et de labs le cas échéant.
• Votre place sera confirmée à la réception d'un devis ou """"booking form""""" signé. Vous recevrez ensuite la convocation et les modalités d'accès en présentiel ou distanciel.
• Attention, si cette formation est éligible au Compte Personnel de Formation, vous devrez respecter un délai minimum et non négociable fixé à 11 jours ouvrés avant le début de la session pour vous inscrire via moncompteformation.gouv.fr.

Accueil des bénéficiaires :

• En cas de handicap : plus d'info sur globalknowledge.fr/handicap
• Le Règlement intérieur est disponible sur globalknowledge.fr/reglement