Google Cloud Platform - Ingénierie de données

Concevoir des architectures de données et des pipelines batch et streaming avec BigQuery, Dataflow et Apache Spark.

A distance

4 jours (28 h) | Prix inter : 3.390,00 € HT | Réf.: CC401

Concevoir une plateforme de données sur **Google Cloud** nécessite de choisir une architecture adaptée et de fiabiliser les traitements. Cette formation associe les architectures lakehouse à la création de **pipelines batch et streaming** avec BigQuery, Dataflow et Apache Spark. Les ateliers permettent de travailler sur l'ingestion, la transformation, la qualité et l'orchestration des données pour relier les choix techniques aux usages professionnels. Cette formation prépare à la certification Google Professional Data Engineer.

A qui s'adresse cette formation ?

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

- Connaître les principes d'ingénierie des données : processus ETL/ELT, modélisation et formats Avro, Parquet et JSON
- Connaître les architectures de lacs de données et d'entrepôts de données
- Maîtriser SQL pour interroger les données. Maîtriser un langage de programmation courant ; Python est recommandé
- Être à l'aise avec les interfaces en ligne de commande
- Connaître les fondamentaux de Google Cloud : calcul, stockage et gestion des identités

Programme

Pendant

1 - Identifier les tâches et composants de l'ingénierie des données sur Google Cloud

- Décrire le rôle de l'ingénieur de données dans la conception des systèmes de traitement
- Distinguer les sources de données des systèmes destinataires
- Identifier les différents formats de données et leurs usages
- Comparer les solutions de stockage disponibles sur Google Cloud
- Identifier les options de gestion des métadonnées
- Expliquer le partage de jeux de données avec Analytics Hub

Charger des données dans BigQuery

2 - Comprendre la réplication et la migration de données sur Google Cloud

- Décrire une architecture de réplication et de migration de données
- Identifier les commandes et les cas d'utilisation de l'outil gcloud
- Comparer les options de déplacement de jeux de données avec Storage Transfer Service et Transfer Appliance
- Expliquer les fonctionnalités de Datastream et ses modalités de déploiement.

3 - Explorer les pipelines d'extraction et de chargement avec BigQuery et BigLake

- Décrire l'architecture d'un pipeline d'extraction et de chargement
- Identifier les options de l'outil en ligne de commande bq
- Expliquer les cas d'utilisation de BigQuery Data Transfer Service
- Examiner l'approche de BigLake pour accéder aux données sans recourir à un modèle d'extraction et de chargement

Prendre en main BigLake

4 - Structurer les pipelines ELT avec BigQuery et Dataform

- Décrire le modèle d'extraction, de chargement et de transformation des données, ou ELT
- Identifier les composants d'un pipeline ELT sur Google Cloud
- Explorer les fonctions de scripting SQL et de planification de BigQuery
- Expliquer le rôle de Dataform dans la gestion des transformations SQL

Créer et exécuter un workflow SQL dans Dataform

5 - Explorer les pipelines ETL avec Dataproc, Dataflow et Bigtable

- Décrire le modèle d'extraction, de transformation et de chargement des données, ou ETL
- Identifier les outils graphiques Google Cloud destinés à la création de pipelines ETL
- Expliquer le traitement des données par lots, ou batch, avec Dataproc
- Examiner l'utilisation de Dataproc Serverless pour Apache Spark dans les traitements ETL
- Comparer les options de traitement des données en continu, ou streaming
- Expliquer la place de Bigtable dans les pipelines de données.

Charger des données dans BigQuery avec Dataproc Serverless pour Apache Spark

Créer un pipeline streaming avec Dataflow pour alimenter un tableau de bord en temps réel

6 - Identifier les techniques d'automatisation des pipelines Google Cloud

- Comparer les modèles d'automatisation des pipelines de données
- Décrire les usages de Cloud Scheduler et de Workflows
- Expliquer le rôle de Cloud Composer dans l'orchestration des traitements
- Identifier les possibilités d'automatisation avec Cloud Run Functions
- Examiner les cas d'utilisation d'Eventarc pour déclencher des traitements

Charger des données dans BigQuery avec Cloud Run Functions

7 - Comparer les architectures data lake, data warehouse et lakehouse

- Distinguer les architectures de lac de données, ou data lake, et d'entrepôt de données, ou data warehouse
- Expliquer les principes de l'approche data lakehouse
- Évaluer les avantages d'une architecture lakehouse
- Choisir une architecture adaptée aux besoins de stockage et d'analyse des données

8 - Construire un lakehouse avec Cloud Storage, Apache Iceberg et BigQuery

- Définir les fondations d'un lac de données avec Cloud Storage.
- Expliquer le rôle du format de table ouvert Apache Iceberg.
- Décrire la place de BigQuery comme moteur central de traitement analytique
- Identifier les usages d'AlloyDB pour les données opérationnelles
- Associer les données opérationnelles et analytiques à l'aide de requêtes fédérées
- Analyser un cas d'utilisation concret d'architecture lakehouse.

9 - Moderniser les entrepôts de données avec BigQuery et BigLake

- Expliquer les fondamentaux de BigQuery et ses capacités d'entreposage de données
- Décrire le partitionnement et le clustering des tables dans BigQuery
- Identifier le rôle de BigLake dans une architecture lakehouse unifiée
- Examiner l'accès aux données externes depuis BigQuery
- Expliquer l'interaction entre BigQuery, BigLake et les tables Apache Iceberg

Interroger des données externes et des tables Apache Iceberg

10 - Analyser les architectures lakehouse avancées et la gouvernance des données

- Identifier les pratiques de gouvernance et de sécurité d'une plateforme de données unifiée
- Examiner les mécanismes de protection des données sensibles et de gestion des métadonnées
- Explorer les usages de l'analytique avancée et de l'apprentissage automatique, ou machine learning, sur un lakehouse
- Analyser des architectures lakehouse réelles et leurs stratégies de migration

Observer les mécanismes de prévention de la perte de données

11 - Identifier les cas d'utilisation des pipelines de données en batch

- Déterminer les situations adaptées au traitement de données par lots
- Décrire le rôle de l'ingénieur de données dans le développement et la maintenance des pipelines batch
- Identifier les étapes du cycle de vie d'un pipeline, de l'ingestion à l'exploitation des données
- Analyser les difficultés liées aux volumes, à la qualité, à la complexité et à la fiabilité des traitements
- Identifier les services Google Cloud permettant de répondre à ces difficultés

12 - Concevoir des pipelines batch évolutifs avec Dataflow et Apache Spark

- Concevoir une architecture de pipeline batch pour l'ingestion de volumes importants
- Définir les transformations de données à réaliser à grande échelle
- Examiner les possibilités de traitement avec Dataflow et Serverless pour Apache Spark
- Identifier les connexions de données et les besoins d'orchestration
- Décrire les étapes d'exécution d'un pipeline Apache Spark
- Optimiser les performances, les ressources et les coûts des traitements batch

Construire un pipeline batch simple avec Serverless pour Apache Spark

Construire un pipeline batch simple avec l'interface Dataflow Job Builder

13 - Contrôler la qualité des données batch avec Apache Spark et Dataflow

- Définir des règles de validation et de nettoyage des données.
- Journaliser les erreurs et analyser leurs causes
- Gérer l'évolution des schémas dans les pipelines batch
- Vérifier l'intégrité des données et identifier les doublons.
- Examiner les techniques de déduplication avec Serverless pour Apache Spark
- Examiner les techniques de déduplication avec Dataflow

Valider la qualité des données d'un pipeline batch avec Serverless pour Apache Spark

14 - Orchestrer et surveiller les pipelines batch avec Cloud Composer

- Décrire les principes d'orchestration des traitements batch
- Examiner la planification des workflows et le suivi du lignage des données
- Identifier le rôle de Cloud Composer dans l'orchestration des pipelines
- Définir une approche d'observabilité unifiée des traitements
- Exploiter les alertes pour identifier et diagnostiquer les incidents
- Examiner la gestion visuelle des pipelines de données

Construire des pipelines batch dans Cloud Data Fusion

15 - Introduire les pipelines streaming et le scénario de mise en pratique

- Identifier les objectifs d'apprentissage et les prérequis du chapitre
- Décrire le rôle des pipelines streaming dans l'ingénierie des données
- Examiner le cas d'utilisation et le contexte de l'entreprise présentés
- Analyser le défi de traitement des données proposé
- Définir la mission à réaliser dans le scénario pédagogique

16 - Analyser les cas d'utilisation du streaming et les architectures de référence

- Décrire les composants d'un pipeline streaming sur Google Cloud
- Identifier les architectures d'extraction, de transformation et de chargement en continu, ou Streaming ETL
- Examiner les usages du streaming pour l'intelligence artificielle et le machine learning
- Distinguer les architectures destinées aux applications de streaming
- Expliquer les cas d'utilisation et les architectures de Reverse ETL

17 - Approfondir Pub/Sub, Apache Kafka, Dataflow, BigQuery et Bigtable

- Comparer les rôles des services Google Cloud dans une architecture streaming
- Définir les concepts de messagerie et les critères de choix entre Pub/Sub et Managed Service for Apache Kafka
- Décrire le traitement des flux avec Dataflow et les difficultés associées aux données en streaming
- Construire et déployer un pipeline de traitement en continu
- Comparer les méthodes d'ingestion des données dans BigQuery
- Examiner les requêtes continues, les traitements ETL et le Reverse ETL avec BigQuery. Configurer un flux de données de Pub/Sub vers BigQuery. Concevoir une architecture de pipeline streaming utilisant BigQuery. Décrire le rôle de Bigtable dans le traitement des données opérationnelles. Établir un pipeline streaming entre Dataflow et Bigtable. Analyser les tendances des flux de données Bigtable avec BigQuery. Synchroniser les résultats d'analyse avec l'application destinée aux utilisateurs

  Traiter des données e-sport avec des pipelines streaming — optionnel

  Enrichir des données de contenus téléchargeables, ou DLC, e-sport avec Apache Beam et Bigtable

  Transmettre des données e-sport en streaming avec Pub/Sub et BigQuery

  Surveiller un chat e-sport avec Streamlit



Les objectifs de la formation

- Comparer les architectures data lake, data warehouse et lakehouse sur Google Cloud
- Concevoir un lakehouse avec Cloud Storage, BigQuery, BigLake et Apache Iceberg
- Créer des workflows SQL de transformation de données avec BigQuery et Dataform
- Développer des pipelines batch avec Dataflow et Apache Spark
- Construire des pipelines streaming avec Pub/Sub, Dataflow et BigQuery
- Appliquer des règles de validation, de nettoyage et de déduplication des données
- Orchestrer et surveiller des pipelines batch avec les services Google Cloud
- Analyser les pratiques de gouvernance et de sécurité des données dans un lakehouse



Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.



Les points forts de la formation

- Une introduction pratique et complète à la conception et au développement de systèmes de traitement de données sur Google Cloud Platform
- Une formation rythmée durant laquelle s'alternent les phases d'apports théoriques, d'échanges, de partages d'expériences et de mises en situation
- Une formation animée par un formateur certifié Google Cloud Platform
- La qualité d'une formation officielle Google (support de cours en anglais)
- 83% des participants à cette formation se sont déclarés satisfaits ou très satisfaits au cours des 12 derniers mois.

Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques.
Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Dates 2027 - Référence CC401

Dernières places disponibles Session garantie

du 18 janv. au 21 janv.	du 18 mai au 21 mai	du 13 sept. au 16 sept.
du 16 févr. au 19 févr.	du 14 juin au 17 juin	du 18 oct. au 21 oct.
du 15 mars au 18 mars	du 19 juil. au 22 juil.	du 15 nov. au 18 nov.
du 19 avr. au 22 avr.	du 17 août au 20 août	du 13 déc. au 16 déc.