[sfeir] Institute

Google Cloud | GCP100DE

Introduction to Data Engineering on Google Cloud

Les compétences essentielles pour comprendre les rôles et responsabilités des Data Engineers ainsi les produits Google Cloud associés

1 jour / 7h

Présentation du cours

Dans ce cours, vous découvrirez l'ingénierie des données sur Google Cloud, les rôles et responsabilités des ingénieurs de données et la manière dont ils s'articulent avec les offres fournies par Google Cloud. Vous découvrirez également comment relever les défis de l'ingénierie des données.

Objectifs pédagogiques

- Comprendre le rôle d'un ingénieur de données.
- Identifier les tâches d'ingénierie des données et les principaux composants utilisés sur Google Cloud.
- Comprendre comment créer et déployer des pipelines de données de différents modèles sur Google Cloud.
- Identifier et utiliser diverses techniques d'automatisation sur Google Cloud.

Public cible

- Ingénieurs de données

- Administrateurs de bases de données
- Administrateurs système

Prérequis

- Expérience préalable de Google Cloud au niveau fondamental en utilisant Cloud Shell et en accédant aux produits à partir de la console Google Cloud.
- Maîtrise de base d'un langage de requête courant tel que SQL.
- Expérience de la modélisation des données et des activités ETL (extraction, transformation, chargement).
- Expérience du développement d'applications à l'aide d'un langage de programmation courant tel que Python.

Programme

Module 01: Tâches et composants de l'ingénierie des données

Sujets :

- Le rôle d'un ingénieur de données
- Sources de données versus puits de données
- Formats de données
- Options de solutions de stockage sur Google Cloud
- Options de gestion des métadonnées sur Google Cloud
- Partage d'ensembles de données à l'aide d'Analytics Hub

Objectifs :

- Expliquez le rôle d'un ingénieur de données.
- Comprenez les différences entre une source de données et un puits de données.
- Expliquez les différents types de formats de données.
- Expliquez les options de solution de stockage sur Google Cloud.
- Découvrez les options de gestion des métadonnées sur Google Cloud.
- Découvrez comment partager facilement des ensembles de données à l'aide d'Analytics Hub.
- Découvrez comment charger des données dans BigQuery à l'aide de la console Google Cloud ou de la gcloud CLI.

Activités :

- Lab : chargement de données dans BigQuery
- Quiz

Module 02 : Réplication et migration des données

Sujets :

- Architecture de réplication et de migration
- L'outil de ligne de commande gcloud
- Déplacement des ensembles de données
- Datastream

Objectifs :

- Expliquez l'architecture de base de réplication et de migration des données de Google Cloud.
- Comprenez les options et les cas d'utilisation de l'outil de ligne de commande gcloud.
- Expliquez les fonctionnalités et les cas d'utilisation de Storage Transfer Service.
- Expliquez les fonctionnalités et les cas d'utilisation de Transfer Appliance.
- Comprenez les fonctionnalités et le déploiement de Datastream.

Activités :

- Lab : Datastream : réplication PostgreSQL vers BigQuery (facultatif pour ILT)
- Quiz

Module 03: Le modèle de pipeline Extract and Load

Sujets :

- Extraire et charger l'architecture
- L'outil de ligne de commande bq
- BigQuery Data Transfer Service
- BigLake

Objectifs :

- Expliquez le diagramme de base d'architecture d'extraction et de chargement.
- Comprendre les options de l'outil de ligne de commande bq.
- Expliquez les fonctionnalités et les cas d'utilisation du service de transfert de données BigQuery.
- Expliquez les fonctionnalités et les cas d'utilisation de BigLake en tant que modèle de chargement sans extraction.

Activities

- Lab : BigLake: Qwik Start
- Quiz

Module 4: Le modèle de pipeline Extract, Load and Transform

Sujets :

- Architecture d'extraction, de chargement et de transformation (ELT)
- Scripting et planification SQL avec BigQuery
- Dataform

Objectifs :

- Expliquer le diagramme de base d'architecture d'extraction, de chargement et de transformation.
- Comprendre un pipeline ELT courant sur Google Cloud.
- Découvrir les fonctionnalités de programmation et de script SQL de BigQuery.
- Expliquer les fonctionnalités et les cas d'utilisation de Dataform.

Activités :

- Lab : Créer et exécuter un workflow SQL dans Dataform
- Quiz

Module 05 : Modèle de pipeline Extract, Transform and Load

Sujets :

- Architecture d'extraction, de transformation et de chargement (ETL)
- Outils d'interface utilisateur graphique Google Cloud pour les pipelines de données ETL
- Traitement de données par lots à l'aide de Dataproc
- Options de traitement de données en continu
- Bigtable et pipelines de données

Objectifs :

- Expliquez le diagramme d'architecture d'extraction, de transformation et de chargement de base.
- Découvrez les outils d'interface utilisateur graphique sur Google Cloud utilisés pour les pipelines de données ETL.
- Expliquez le traitement de données par lots à l'aide de Dataproc.
- Apprenez à utiliser Dataproc Serverless pour Spark pour ETL.

- Expliquez les options de traitement de données en streaming.
- Expliquez le rôle que joue Bigtable dans les pipelines de données.

Activités :

- Lab : Utiliser Dataproc Serverless pour Spark pour charger BigQuery (facultatif pour ILT)
- Lab : Créer un pipeline de données en continu pour un tableau de bord en temps réel avec Dataflow
- Quiz

Module 06: Techniques d'automatisation

Sujets :

- Modèles et options d'automatisation pour les pipelines
- Cloud Scheduler et Workflows
- Cloud Composer
- Cloud Run Functions
- Eventarc

Objectifs :

- Expliquez les modèles et options d'automatisation disponibles pour les pipelines.
- Découvrez Cloud Scheduler et Workflows.
- Découvrez Cloud Composer.
- Découvrez les fonctions de Cloud Run.
- Expliquez les cas d'utilisation des fonctionnalités et de l'automatisation pour Eventarc

Activités :

- Lab : Utiliser les fonctions Cloud Run pour charger BigQuery (facultatif pour ILT)
- Quiz