GOOGLE CLOUD PLATFORM INGÉNIERIE DE DONNÉES

Durée	4 jours	Référence Formation	4-GPGU

Objectifs

Apprendre à concevoir et déployer des pipelines et des architectures pour le traitement des données
Comprendre comment créer et déployer des workflows de machine learning
Être capable d'interroger des ensembles de données
Comprendre comment visualiser des résultats des requêtes et créer des rapports

Participants

Développeurs expérimentés en charge des transformations du Big Data

Pré-requis

Maîtriser les principes de base des langages de requête courants tels que SQL Avoir de l'expérience en modélisation, extraction, transformation et chargement des données Savoir développer des applications à l'aide d'un langage de programmation courant tel que Python Savoir utiliser le Machine Learning et/ou les statistiques

Moyens pédagogiques

Accueil des stagiaires dans une salle dédiée à la formation équipée d'un vidéo projecteur, tableau blanc et paperboard ainsi qu'un ordinateur par participant pour les formations informatiques.
Positionnement préalable oral ou écrit sous forme de tests d'évaluation, feuille de présence signée en demi-journée, évaluation des acquis tout au long de la formation.
En fin de stage : QCM, exercices pratiques ou mises en situation professionnelle, questionnaire de satisfaction, attestation de stage, support de cours remis à chaque participant.
Formateur expert dans son domaine d'intervention
Apports théoriques et exercices pratiques du formateur
Utilisation de cas concrets issus de l'expérience professionnelle des participants
Réflexion de groupe et travail d'échanges avec les participants
Pour les formations à distance : Classe virtuelle organisée principalement avec l'outil ZOOM.
Assistance technique et pédagogique : envoi des coordonnées du formateur par mail avant le début de la formation pour accompagner le bénéficiaire dans le déroulement de son parcours à distance.

PROGRAMME

- **Introduction à l'ingénierie des données**

Explorer le rôle d'un data engineer
Analyser les défis d'ingénierie des données
Introduction à BigQuery
Data lakes et data warehouses
Démo: requêtes fédérées avec BigQuery
Bases de données transactionnelles vs data warehouses
Démo: recherche de données personnelles dans votre jeu de données avec l'API DLP
Travailler efficacement avec d'autres équipes de données
Gérer l'accès aux données et gouvernance
Construire des pipelines prêts pour la production

CAP ÉLAN FORMATION
www.capelanformation.fr - Tél : 04.86.01.20.50
Mail : contact@capelanformation.fr
Organisme enregistré sous le N° 76 34 0908834
[version 2023]

Etude de cas d'un client GCP
Lab : Analyse de données avec BigQuery

- **Construire un Data lake**
Introduction aux data lakes
Stockage de données et options ETL sur GCP
Construction d'un data lake à l'aide de Cloud Storage
Démo : optimisation des coûts avec les classes et les fonctions cloud de Google Cloud Storage
Sécurisation de Cloud Storage
Stocker tous les types de données
Démo : exécution de requêtes fédérées sur des fichiers Parquet et ORC dans BigQuery
Cloud SQL en tant que data lake relationnel

- **Construire un Data Warehouse**
Le data warehouse moderne
Introduction à BigQuery
Démo : Requêter des TB + de données en quelques secondes
Commencer à charger des données
Démo: Interroger Cloud SQL à partir de BigQuery
Lab : Chargement de données avec la console et la CLI
Explorer les schémas
Exploration des jeux de données publics BigQuery avec SQL à l'aide de Information_Schema
Conception de schéma
Démo : Exploration des jeux de données publics BigQuery avec SQL à l'aide de Information_Schema
Champs imbriqués et répétés dans BigQuery
Lab : tableaux et structures
Optimiser avec le partitionnement et le clustering
Démo : Tables partitionnées et groupées dans BigQuery
Aperçu : Transformation de données par lots et en continu

- **Introduction à la construction de pipelines de données par lots EL, ELT, ETL**
Considérations de qualité
Comment effectuer des opérations dans BigQuery
Démo : ETL pour améliorer la qualité des données dans BigQuery
Des lacunes
ETL pour résoudre les problèmes de qualité des données

- **Exécution de Spark sur Cloud Dataproc**
L'écosystème Hadoop
Exécution de Hadoop sur Cloud Dataproc GCS au lieu de HDFS
Optimiser Dataproc
Atelier : Exécution de jobs Apache Spark sur Cloud Dataproc

- **Traitement de données sans serveur avec Cloud dataflow**
Cloud Dataflow
Pourquoi les clients apprécient-ils Dataflow ?
Pipelines de flux de données
Lab : Pipeline de flux de données simple (Python / Java)
Lab : MapReduce dans un flux de données (Python / Java)
Lab : Entrées latérales (Python / Java)

Templates Dataflow
Dataflow SQL

- **Gestion des pipelines de données avec Cloud Data fusion and Cloud composer**
Création visuelle de pipelines de données par lots avec Cloud Data Fusion: composants, présentation de l'interface utilisateur, construire un pipeline, exploration de données en utilisant Wrangler
Lab : Construction et exécution d'un graphe de pipeline dans Cloud Data Fusion
Orchestrer le travail entre les services GCP avec Cloud Composer - Apache Airflow
Environment : DAG et opérateurs, planification du flux de travail
Démo : Chargement de données déclenché par un événement avec Cloud Composer, Cloud Functions, Cloud Storage et BigQuery
Lab : Introduction à Cloud Composer

- **Introduction au traitement de données en streaming**
Traitement des données en streaming

- **Serverless messaging avec Cloud Pub/Sub**
Cloud Pub/Sub
Lab : Publier des données en continu dans Pub/Sub

- **Fonctionnalités streaming de Cloud Dataflow**
Fonctionnalités streaming de Cloud Dataflow
Lab : Pipelines de données en continu

- **Fonctionnalités streaming à haut débit BIGQUERY ET BIGTABLE**
Fonctionnalités de streaming BigQuery
Lab : Analyse en continu et tableaux de bord
Cloud Bigtable
Lab : Pipelines de données en continu vers Bigtable

- **Fonctionnalités avancées de BIGQUERY et performance**
Analytic Window Functions
Utiliser des clauses With
Fonctions SIG
Démo: Cartographie des codes postaux à la croissance la plus rapide avec BigQuery GeoViz
Considérations de performance
Lab : Optimisation de vos requêtes BigQuery pour la performance
Lab : Création de tables partitionnées par date dans BigQuery

- **Introduction à l'analytique et à l'IA**
Qu'est-ce que l'IA?
De l'analyse de données ad hoc aux décisions basées sur les données
Options pour modèles ML sur GCP

- **API de modèle ML prédéfinis pour les données non structurées**
Les données non structurées sont difficiles à utiliser
API ML pour enrichir les données
Lab : Utilisation de l'API en langage naturel pour classer le texte non structuré

- **Big Data Analytics avec les notebooks Cloud AI plateform**

Qu'est-ce qu'un notebook
BigQuery Magic et liens avec Pandas
Lab : BigQuery dans Jupyter Labs sur IA Platform

- **Pipeline de production ML avec Kubeflow**
Façons de faire du ML sur GCP
Kubeflow AI Hub
Lab : Utiliser des modèles d'IA sur Kubeflow

- **Création de modèles personnalisés avec SQL dans BIGQUERY ML**
BigQuery ML pour la construction de modèles rapides
Démo : Entraîner un modèle avec BigQuery ML pour prédire les tarifs de taxi à New York
Modèles pris en charge
Lab : Prédire la durée d'une sortie à vélo avec un modèle de régression dans BigQuery ML
Lab : Recommandations de film dans BigQuery ML

- **Création de modèles personnalisés avec Cloud AUTOML**
Pourquoi Auto ML?
Auto ML Vision
Auto ML NLP
Auto ML Tables