Google Cloud Platform
Ingénierie de données

Mise à jour nov. 2023

Durée 4 jours (28 heures)

« Délai d'accès maximum 1 mois »

Infrastructure

OBJECTIFS PROFESSIONNELS

- 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

PRE-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 PEDAGOGIQUES

- Réflexion de groupe et apports théoriques du formateur
- Travail d'échange avec les participants sous forme de
- Utilisation de cas concrets issus de l'expérience professionnelle
- Validation des acquis par des questionnaires, des tests d'évaluation, des mises en situation et des jeux pédagogiques.
- Remise d'un support de cours.

MODALITES D'EVALUATION

- Feuille de présence signée en demi-journée,
- Evaluation des acquis tout au long de la formation,
- Questionnaire de satisfaction,
- Positionnement préalable oral ou écrit,
- Evaluation formative tout au long de la formation,
- Evaluation sommative faite par le formateur ou à l'aide des certifications disponibles,
- Sanction finale : Certificat de réalisation, certification éligible au RS selon l'obtention du résultat par le stagiaire

MOYENS TECHNIQUES EN PRESENTIEL

- Accueil des stagiaires dans une salle dédiée à la formation, équipée d'ordinateurs, d'un vidéo projecteur d'un tableau blanc et de paperboard. Nous préconisons 8 personnes maximum par action de formation en présentiel

MOYENS TECHNIQUES DES CLASSES EN CAS DE FORMATION DISTANCIELLE

- A l'aide d'un logiciel comme Teams, Zoom etc... un micro et éventuellement une caméra pour l'apprenant,
- suivez une formation uniquement synchrone en temps réel et entièrement à distance. Lors de la classe en ligne, les apprenants interagissent et comuniquent entre eux et avec le formateur.
- Les formations en distanciel sont organisées en Inter-Entreprise comme en Intra-Entreprise.
- L'accès à l'environnement d'apprentissage (support de cours, labs) ainsi qu'aux preuves de suivi et d'assiduité (émargement, évaluation) est assuré. Nous préconisons 4 personnes maximum par action de formation en classe à distance

ORGANISATION

- Les cours ont lieu de 9h à 12h30 et de 14h à 17h30.

PROFIL FORMATEUR

- Nos formateurs sont des experts dans leurs domaines d'intervention

- Leur expérience de terrain et leurs qualités pédagogiques constituent un gage de qualité.

A L'ATTENTION DES PERSONNES EN SITUATION DE HANDICAP

- Les personnes atteintes de handicap souhaitant suivre cette formation sont invitées à nous contacter directement, afin d'étudier ensemble les possibilités de suivre la formation.

Programme de formation

Introduction à l'ingénierie des données (03h45)

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

Construire un Data lake (02h15)

- 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 (04h15)

- 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 (01h30)

- 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 (01h15)

- 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 (02h15)

- 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 (01h45)

- 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 (00h15)

- Traitement des données en streaming

Serverless messaging avec Cloud Pub/Sub (00h30)

- Cloud Pub/Sub
- Lab : Publier des données en continu dans Pub/Sub

Fonctionnalités streaming de Cloud Dataflow (00h30)

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

Fonctionnalités streaming à haut débit BIGQUERY ET BIGTABLE (01h15)

- 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 (02h00)

- 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 (00h45)

- 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 (00h45)

- 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 (00h45)

- 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 (00h45)

- 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 (01h30)

- 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 (01h15)

- Pourquoi Auto ML?
- Auto ML Vision
- Auto ML NLP
- Auto ML Tables