SARL ONEO
1025 rue Henri Becquerel
Parc Club du Millénaire, Bât. 27
34000 MONTPELLIER
N° organisme : 91 34 0570434

Nous contacter :
+33 (0)4 67 13 45 45
www.ait.fr
contact@ait.fr

Référence	**4-GO-INDO**
Durée	**4 jours (28 heures)**
Éligible CPF	**NON**
Mise à jour	**15/12/2022**

Google Cloud Platform Ingénierie de données

Objectifs pedagogiques

- 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

Public concerne

- Développeurs expérimentés en charge des transformations du Big Data

Prerequis

- 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 réunion-discussion
- 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
- Alternance entre apports théoriques et exercices pratiques (en moyenne 30 à 50%)
- 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,
- Attestation de stage à chaque apprenant,
- 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

Moyens techniques en presentiel

- Accueil des stagiaires dans une salle dédiée à la formation équipée à minima d'un vidéo projecteur et d'un tableau blanc et/ou paperboard.
- Pour les formations nécessitant un ordinateur, un PC est mis à disposition de chaque participant.

Moyens techniques en distanciel

- A l'aide d'un logiciel (Teams, Zoom...), d'un micro et éventuellement d'une caméra les apprenants interagissent et communiquent entre eux et avec le formateur.
- Sessions organisées en inter comme en intra entreprise.
- L'accès à l'environnement d'apprentissage ainsi qu'aux preuves de suivi et d'assiduité (émargement, évaluation) est assuré.
- Pour toute question avant et pendant le parcours, assistance technique à disposition au 04 67 13 45 45.

Organisation

- Délai d'accès : 5 jours ouvrés
(délai variable en fonction du financeur)
- Les cours ont lieu de 9h à 12h30 et de 13h30 à 17h

Accessibilite

- Les personnes en situation d'handicap sont invitées à nous contacter directement, afin d'étudier ensemble les possibilités de suivre la formation.
- Pour tout renseignement, notre référent handicap reste à votre disposition : mteyssedou@ait.fr

Profil formateur

- Formateur expert du domaine.
- Leur expérience de terrain et leurs qualités pédagogiques constituent un gage de qualité.

Certification possible

- Aucune

Google Cloud Platform Ingénierie de données

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
- ☐ 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 ET 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

FONCTIONNALITÉ 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 PRÉDÉFINIES 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 PLATFORM

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

PIPELINES 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