Data Engineering on Google Cloud Platform
Cours officiel, préparation aux examens de certification Google Cloud

Cours Pratique de 4 jours - 28h
Réf : DGC - Prix 2024 : 3 730€ HT
Le prix pour les dates de sessions 2025 pourra être révisé

Avec cette formation, vous apprendrez à concevoir et créer des systèmes de traitement des données sur Google Cloud Platform. Grâce à de nombreux travaux pratiques, vous apprendrez à concevoir des systèmes de traitement des données, à construire des pipelines de données de bout en bout, à analyser les données et à effectuer un apprentissage automatique. Cette formation couvre les données structurées, non structurées et en streaming.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Concevoir et développer des systèmes de traitement des données sur Google Cloud

Traiter des données par lot ou par flux en mettant en œuvre des pipelines de données d'autoscaling sur Dataflow

Obtenir des insights métier à partir d'ensembles de données extrêmement volumineux à l'aide de BigQuery

Exploiter des données non structurées à l'aide de Spark et des interfaces de programmation de ML sur Dataproc

Obtenir des insights immédiats à partir de flux de données

Découvrir les API de machine learning (ML) et BigQuery ML, et apprendre à utiliser Cloud AutoML

MÉTHODES PÉDAGOGIQUES

Animation de la formation en français. Support de cours officiel en anglais.

CERTIFICATION

Nous vous recommandons de suivre cette formation si vous souhaitez préparer la certification "Google Cloud Professional Data Engineer".

LE PROGRAMME

dernière mise à jour : 09/2021

1) Introduction à l'ingénierie des données

- Explorer le rôle d'un data engineer.
- Analyser les défis de l'ingénierie des données.
- Introduction à BigQuery.
- Les data lakes et les data warehouses.
- Démonstration "Federated Queries avec BigQuery".
- Bases de données transactionnelles versus data warehouses.
- Démonstration "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.

PARTICIPANTS

Développeurs expérimentés responsables de la gestion des transformations des méga données notamment l'extraction, le chargement, la transformation, le nettoyage et la validation des données.

PRÉREQUIS

Avoir suivi "Google Cloud Fundamentals : big data et machine learning" ou connaissances équivalentes et des compétences en langage de requête, en modélisation de données, en Python et en statistiques.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils sont agréés par l'éditeur et sont certifiés sur le cours. Ils ont aussi été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum trois à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Évaluation des compétences visées en amont de la formation.
Évaluation par le participant, à l'issue de la formation, des compétences acquises durant la formation.
Validation par le formateur des acquis du participant en précisant les outils utilisés : QCM, mises en situation…
À l'issue de chaque stage, ITTCERT fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques. Les participants réalisent aussi une évaluation officielle de l'éditeur.
Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le participant a bien assisté à la totalité de la session.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

Les ressources pédagogiques utilisées sont les supports et les travaux pratiques officiels de l'éditeur.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

- Étude de cas d'un client Google Cloud Platform (GCP).
Travaux pratiques : Analyse de données avec BigQuery.

2) 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.
- Démonstration : optimisation des coûts avec les classes et les fonctions cloud de Google Cloud Storage.
- Sécurisation de Cloud Storage.
- Stocker tous les types de données.
- Démonstration : exécution de requêtes fédérées sur des fichiers Parquet et ORC dans BigQuery.
- Cloud SQL en tant que data lake relationnel.
Travaux pratiques : Charger la BDD Taxis dans le Cloud SQL.

3) Construire un data warehouse

- Le data warehouse moderne.
- Introduction à BigQuery.
- Démonstration : requêtes de Terabits de données en quelques secondes.
- Chargement de données.
- Démonstration : interroger Cloud SQL à partir de BigQuery.
- Explorer les schémas.
- Exploration des jeux de données publics BigQuery avec SQL à l'aide de INFORMATION_SCHEMA.
- Conception de schémas.
- Champs imbriqués et répétés.
- Champs imbriqués et répétés dans BigQuery.
- Optimiser le partitionnement et le clustering.
- Démonstration : tables partitionnées et groupées dans BigQuery.
- Transformation de données par lots et en continu.
Travaux pratiques : Charger des données avec la console et la CLI. Travailler avec les tableaux et les structures.

4) Introduction à la construction de pipelines de données par lots

- Les approches d'intégration EL, ELT et ETL (Extraction, chargement et transformation de données).
- Les considérations de qualité.
- Comment effectuer des opérations dans BigQuery.
- Démonstration : ELT pour améliorer la qualité des données dans BigQuery.
- Les lacunes.
- ETL pour résoudre les problèmes de qualité.

5) Exécution de Spark sur Cloud Dataproc

- L'écosystème Hadoop.
- Exécution de Hadoop sur Cloud Dataproc GCS au lieu de HDFS.
- Optimiser Dataproc.
Travaux pratiques : Exécuter des jobs Apache Spark sur Cloud Dataproc.

6) Traitement de données sans serveur avec Cloud Dataflow

- Cloud Dataflow.
- Pourquoi les clients apprécient-ils Dataflow ?
- Pipelines de flux de données.
- Templates Dataflow.
- Dataflow SQL.
Travaux pratiques : Pipeline de flux de données simple (Python/Java). MapReduce dans un flux de données (Python/Java). Entrées latérales (Python/Java).

7) Gestion des pipelines de données avec Cloud Data Fusion et Cloud Composer

- Création visuelle de pipelines de données par lots avec Cloud Data Fusion.
- Orchestrer le travail entre les services GCP avec Cloud Composer - Apache Airflow Environnement - DAG et opérateurs.
- Démonstration : chargement de données déclenché par un événement avec Cloud Composer, Cloud Functions, Cloud Storage...
- Surveillance et journalisation.

Travaux pratiques : Construire et exécuter un graphe de pipeline dans Cloud Data Fusion (composants, présentation de l'interface utilisateur, construire un pipeline, exploration de données en utilisant Wrangler). Utilisation de Cloud Composer.

8) Introduction au traitement de données en streaming

- Traitement des données en streaming.

9) Serverless messaging avec Cloud Pub/Sub

- Présentation de Cloud Pub/Sub.

Travaux pratiques : Publier des données en continu dans Pub/Sub.

10) Fonctionnalités streaming du Cloud Dataflow

- Fonctionnalités streaming de Cloud Dataflow.

Travaux pratiques : Pipelines de données en continu.

11) Fonctionnalités streaming à haut débit BigQuery et Bigtable

- Fonctionnalités streaming BigQuery.
- Cloud Bigtable.

Travaux pratiques : Analyse en continu et tableaux de bord. Pipelines de données en continu vers Bigtable.

12) Fonctionnalités avancées de BigQuery et performance

- Fonctionnalités "Analytic Window".
- Utilisation des clauses With.
- Fonctions SIG.
- Démonstration : cartographie des codes postaux à la croissance la plus rapide avec BigQuery GeoViz.
- Considérations de performance.

Travaux pratiques : Optimiser vos requêtes BigQuery pour la performance. Créer des tables partitionnées par date dans BigQuery (optionnel).

13) Introduction à l'analytique et à l'intelligence artificielle

- Qu'est-ce que l'intelligence artificielle (IA) ?
- De l’analyse de données ad hoc aux décisions basées sur les données.
- Options pour modèles de machine learning (ML) sur Google Cloud Platform.

14) API de modèles de ML prédéfinies pour les données non structurées

- Les données non structurées sont difficiles à utiliser.
- API ML pour enrichir les données.

Travaux pratiques : Utiliser l’interface de programmation des applications (API) en langage naturel pour classer le texte non structuré.

15) Big Data Analytics avec les notebooks Cloud AI Platform

- Qu'est-ce qu'un notebook ?
- BigQuery Magic et liens avec Pandas.

Travaux pratiques : BigQuery dans Jupyter Labs sur IA Platform.

16) Pipelines de production de machine learning avec Kubeflow

- Façons de faire du machine learning (ML) sur Google Cloud Platform.
- Kubeflow AI Hub.

- Artificial Intelligence (AI) Hub.
Travaux pratiques : Utiliser des modèles d'IA sur Kubeflow.

17) Création de modèles personnalisés avec SQL dans BigQuery ML

- BigQuery ML pour la construction de modèles rapides.
- Démonstration : entraîner un modèle avec BigQuery ML pour prédire les tarifs de taxis à New York.
- Modèles pris en charge.
Recommandations de films dans BigQuery ML.

18) Création de modèles personnalisés avec Cloud AutoML

- Pourquoi AutoML ?
- Auto ML Vision.
- Auto ML Natural Language Processing (NLP).
- Auto ML Tables.

LES DATES

CLASSE À DISTANCE	PARIS
2024 : 16 déc.	2024 : 16 déc.
2025 : 28 janv., 14 avr., 24 juin, 22 sept.	2025 : 28 janv., 14 avr., 24 juin, 22 sept.