

Mis à jour le 03/05/2024

S'inscrire

Formation DataProc

2 jours (14 heures)

Présentation

Notre formation DataProc vous permettra de réaliser de la manipulation de données complexes pour le traitement par lot, l'émission de requêtes, le streaming et le machine learning. DataProc est un service géré Hadoop et Spark qui vous permet de créer des cluster de données extrêmement rapidement et de les gérer à moindre coût.

Notre programme vous apprendra la gestion des données sur GCP, notamment les concepts de big data et les solutions disponibles sur Google Cloud Platform. Vous serez en mesure d'utiliser le tableau de bord Cloud Dataproc pour la création de projets.

Cette formation vous apprendra également les processus de création et de gestion des clusters de données. Pour une utilisation optimale de l'outil, les notions d'infrastructure as Code ([IaC](#)) ainsi que l'utilisation avec Terraform seront aussi abordés.

Comme toutes nos formations, elle se déroulera sur la dernière version de l'outil : [Dataproc 2.2](#)

Objectifs

- Comprendre la gestion de données avec DataProc
- Utiliser l'outil pour la création et la gestion de cluster
- Utiliser DataProc dans un contexte d'infrastructure as Code

Public visé

- Data Scientists
- Ingénieurs IA

Pré-requis

- Connaissance de Google Cloud Platform
- Connaissance de Terraform

PROGRAMME DE NOTRE FORMATION DATAPROC

INTRODUCTION À DATAPROC ET AU BIG DATA

- Présentation des concepts de big data et des solutions
- Définition de Cloud Dataproc et de ses avantages par rapport aux solutions traditionnelles
- Méthodes d'accès à Dataproc
- Navigation dans le tableau de bord
- Création de projets et gestion des ressources dans Dataproc

CRÉATION ET GESTION DES CLUSTERS

- Processus détaillé de création d'un cluster Dataproc
- Suppression et gestion du cycle de vie d'un cluster
- Rôles des nœuds master et worker
- Personnalisation des clusters avec des types de machines et des workers préemptibles
- Gestion des identités et des accès, permissions et rôles

INTÉGRATION AVEC D'AUTRES SERVICES GCP

- Utilisation avec BigQuery pour l'analyse interactive de données
- Stockage et gestion des données avec Cloud Storage
- Intégration avec les services de bases de données Cloud SQL et Firestore
- Automatisation des workflows avec les modèles de workflow Dataproc

TRAITEMENT DES DONNÉES

- Exécution de traitements par lots (batch) et en temps réel (streaming)
- Écriture, soumission et gestion des jobs Hadoop et Spark
- Surveillance, journalisation et débogage des jobs et des clusters
- Autoscaling et optimisation des performances des clusters

INFRASTRUCTURE AS CODE AVEC TERRAFORM SUR GCP

- Introduction à l'Infrastructure as Code et aux avantages de Terraform
- Gestion de l'infrastructure GCP de manière déclarative avec Terraform
- Écriture et organisation de code Terraform pour les ressources Dataproc
- Bonnes pratiques pour la modularité et la réutilisabilité du code Terraform

FONCTIONNALITÉS AVANCÉES ET OPTIMISATION

- Personnalisation des clusters avec des actions d'initialisation et des images personnalisées
- Techniques d'optimisation de performance pour le traitements de données
- Stratégies de haute disponibilité et de tolérance aux pannes pour les clusters
- Conseils pour optimiser les coûts des clusters à long terme

Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.