Serverless Data Processing with Dataflow
Cours officiel, préparation aux examens de certification Google Cloud

Cours Pratique de 3 jours - 21h
Réf : SDD - Prix 2024 : 2 790€ HT

Avec cette formation, vous découvrirez comment Apache Beam et Dataflow fonctionnent ensemble pour répondre à vos besoins de traitement de données sans risque de dépendance vis-à-vis d'un fournisseur. Vous apprendrez comment convertir votre logique métier en applications de traitement de données pouvant s'exécuter sur Dataflow. La formation se termine par un focus sur les opérations qui passe en revue les leçons les plus importantes pour exploiter une application de données sur Dataflow, y compris la surveillance, le dépannage, les tests et la fiabilité.

OBJECTIFS PÉDAGOGIQUES

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

Démontrer comment Apache Beam et Dataflow fonctionnent ensemble

Résumer les avantages de Beam Portability Framework et l'activer pour vos pipelines Dataflow.

Activer Shuffle et Streaming Engine, pour les pipelines batch et streaming, pour des performances maximales

Activer la planification flexible des ressources pour des performances plus rentables

Sélectionner la bonne combinaison d'autorisations IAM pour votre tâche Dataflow

Mettre en œuvre les meilleures pratiques pour un environnement de traitement de données sécurisé

Sélectionner et ajuster les E/S de votre choix pour votre pipeline Dataflow

Utiliser des schémas pour simplifier votre code Beam et améliorer les performances de votre pipeline

Développer un pipeline Beam en utilisant SQL et DataFrames

Effectuer la surveillance, le dépannage, les tests et la CI/CD sur les pipelines Dataflow

MÉTHODES PÉDAGOGIQUES

Animation de la formation en français. Support de cours officiel au format numérique et en anglais. Bonne compréhension de l'anglais à l'écrit.

CERTIFICATION

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

PARTICIPANTS

Data engineer, data analysts et data scientists aspirant à développer des compétences en ingénierie des données.

PRÉREQUIS

Avoir suivi le cours "Data Engineering on Google Cloud Platform" Réf DGC ou avoir des connaissances équivalentes.

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.

LE PROGRAMME

dernière mise à jour : 10/2023

1) Portabilité de Beam

- Résumer les avantages du Beam Portability Framework.
- Personnaliser l'environnement de traitement des données de votre pipeline à l'aide de conteneurs personnalisés.
- Examiner les cas d'utilisation pour les transformations Cross-Language.

- Activer le Beam Portability Framework pour vos pipelines Dataflow.

2) Séparer le calcul et le stockage avec Dataflow

- Activer Shuffle et Streaming Engine, pour les pipelines batch et streaming, pour des performances maximales.
- Activer la planification flexible des ressources pour des performances plus rentables.

3) IAM, Quotas et Permissions

- Sélectionner la bonne combinaison d'autorisations IAM pour votre tâche Dataflow.
- Déterminer vos besoins en capacité en inspectant les quotas pertinents pour vos tâches Dataflow.

4) Sécurité

- Sélectionner une stratégie de traitement des données zonales à l'aide de Dataflow.
- Mettre en œuvre les meilleures pratiques pour un environnement de traitement de données sécurisées.

5) Revue des concepts de Beam

- Passer en revue les principaux concepts d'Apache Beam (Pipeline, PCollections, PTransforms, Runner, lecture/écriture..).
- Passer en revue les bundles et le cycle de vie DoFn.

6) Windows, Watermarks, Triggers

- Implémenter une logique pour gérer vos données tardives.
- Passer en revue les différents types de déclencheurs.
- Passer en revue les principaux concepts de diffusion en continu (unbounded PCollections, windows).

7) Sources et Sinks

- Écrire sur les IO de votre choix pour votre pipeline Dataflow.
- Ajuster votre transformation Source/Sink pour des performances maximales.
- Créer des Sources et des sinks personnalisés à l'aide de SDF.

8) Schémas

- Introduire des schémas qui donnent aux développeurs un moyen d'exprimer des données dans leurs pipelines Beam.
- Utiliser des schémas pour simplifier votre code Beam et améliorer les performances de votre pipeline.

9) État et Timers

- Identifier les cas d'utilisation pour les implémentations d'API d'état et de timer.
- Sélectionner le bon type d'état et de timers pour votre pipeline.

10) Bonnes pratiques

- Mettre en œuvre les bonnes pratiques pour les pipelines Dataflow.

11) Dataflow SQL et DataFrames

- Développer un pipeline Beam en utilisant SQL et DataFrames.

12) Notebooks Beam

- Prototyper votre pipeline en Python à l'aide des notebooks Beam.
- Lancer une tâche dans Dataflow à partir d'un notebooks.

13) Monitoring

- Accéder à l'interface utilisateur des détails de la tâche Dataflow.
- Interpréter les graphiques de métriques de travail pour diagnostiquer les régressions du pipeline.
- Définir des alertes sur les tâches Dataflow à l'aide de Cloud Monitoring.

- Utiliser les journaux Dataflow et les widgets de diagnostic pour résoudre les problèmes de pipeline.

14) Dépannage et débogage

- Utiliser une approche structurée pour déboguer vos pipelines Dataflow.
- Examiner les causes courantes des défaillances de pipeline.

15) Performance

- Comprendre les considérations de performances pour les pipelines.
- Tenir compte de la façon dont la forme de vos données peut affecter les performances du pipeline.

16) Testing et CI/CD

- Approches de test pour votre pipeline Dataflow.
- Passez en revue les frameworks et les fonctionnalités disponibles pour rationaliser votre flux de travail CI/CD.

17) Fiabilité

- Mettre en œuvre les bonnes pratiques en matière de fiabilité pour vos pipelines Dataflow.

18) Flex Templates

- Utiliser des Flex Templates pour standardiser et réutiliser le code du pipeline Dataflow.

LES DATES

CLASSE À DISTANCE
2024 : 10 sept.