[sfeir] Institute

Google Cloud | GCP300DATAFLOW

Serverless Data Processing with Dataflow

Approfondissez votre maîtrise de Dataflow

3jours / 21h

Objectifs pédagogiques

- Démontrer comment Apache Beam et Dataflow fonctionnent ensemble pour répondre aux besoins de traitement des données de votre organisation.
- Résumer les avantages de Beam Portability Framework et activer-le pour vos pipelines Dataflow.
- Activer Shuffle et Streaming Engine, respectivement 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.

Modalités d'évaluation : Les objectifs pédagogiques sont évalués à travers la réalisation des parties pratiques (labs dirigés) sous la supervision du formateur délivrant la session de formation.

Public cible

- Data Engineer
- Data Analysts et Data Scientists aspirant à développer des compétences en ingénierie des données

Prérequis

Pour tirer le meilleur parti de ce cours, les participants doivent :

- Avoir suivi le modules « Créer des pipelines de données par lots » dans le cours Data Engineering on Google Cloud ou avoir des connaissances équivalentes
- Avoir suivi le module « Créer des systèmes d'analyse de flux résilients » dans le cours Data Engineering on Google Cloud ou avoir des connaissances équivalentes

Programme

Module 1: Introduction

Sujets

- Présentation du cours
- Actualisation des faisceaux et des flux de données

Objectifs

- Présentation des objectifs du cours.
- Démontrer comment Apache Beam et Dataflow fonctionnent ensemble pour répondre aux besoins de traitement des données de votre organisation.

Module 2: Portabilité de Beam

Sujets

- Portabilité de Beam
- Runner v2
- Environnements de conteneurs
- Transformations Cross-Language

Objectifs

- 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.
- Activez le Beam Portability Framework pour vos pipelines Dataflow.

Activités

- Quiz

Module 3: Séparer le calcul et le stockage avec Dataflow

Sujets

- Dataflow
- Dataflow Shuffle Service
- Dataflow Streaming Engine
- Flexible Resource Scheduling

Objectifs

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

Activités

- Quiz

Module 4: IAM, Quotas et Permissions

Sujets

- IAM
- Quota

Objectifs

- 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.

Activités

- Quiz

Module 5: Sécurité

Sujets

- Localité des données
- Shared VPC
- IPs privées
- CMEK

Objectifs

- Sélectionner votre stratégie de traitement des données zonales à l'aide de Dataflow, en fonction de vos besoins en matière de localisation des données.
- Mettre en œuvre les meilleures pratiques pour un environnement de traitement de données sécurisé.

Activités

- Lab pratique et quiz

Module 6: Revue des concepts de BEAM

Sujets

- Les bases Beam
- Transformations utilitaires
- Cycle de vie DoFn

Objectifs

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

Activités

- Lab pratique et quiz

Module 7: Windows, Watermarks, Triggers

Sujets

- Windows
- Watermarks
- Triggers

Objectifs

- 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).

Activités

- Lab pratique et quiz

Module 8: Sources and Sinks

Sujets

- Sources et Sinks
- Text IO et File IO
- BigQuery IO
- PubSub IO
- Kafka IO
- Bigable IO
- Avro IO
- Splittable DoFn

Objectifs

- 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.

Activités

- Quiz

Module 9: Schémas

Sujets

- Beam Schemas
- Exemples de code

Objectifs

- Introduire des schémas, qui donnent aux développeurs un moyen d'exprimer des données structurées dans leurs pipelines Beam.

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

Activités

- Lab pratique et quiz

Module 10: État et Timers

Sujets

- State API
- Timer API
- Summary

Objectifs

- 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

Activités

- Quiz

Module 11: Bonnes pratiques

Sujets

- Schémas
- Gestion des données non traitables
- La gestion des erreurs
- Générateur de code AutoValue
- Traitement des données JSON
- Utiliser le cycle de vie DoFn
- Optimisations de pipeline

Objectifs

- Implement best practices for Dataflow pipelines.

Activités

- Lab pratique et quiz

Module 12: Dataflow SQL et DataFrames

Sujets

- Dataflow et Beam SQL
- Windowing in SQL
- Beam DataFrames

Objectifs

- Développer un pipeline Beam en utilisant SQL et DataFrames.

Activités

- Lab pratique et quiz

Module 13: Beam Notebooks

Sujets

- Beam Notebooks

Objectifs

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

Activités

- Quiz

Module 14: Monitoring

Sujets

- Job List
- Job Info
- Job Graph
- Job Metrics
- Metrics Explorer

Objectifs

- 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.

Activités

- Quiz

Module 15: Monitoring

Sujets

- Logging
- Rapport d’erreur

Objectifs

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

Activités

- Quiz

Module 16: Dépannage et débogage

Sujets

- Flux de travail de dépannage
- Types de problèmes

Objectifs

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

Activités

- Lab pratique et quiz

Module 17: Performance

Sujets

- Conception de pipelines
- Forme des données
- Source, Sinks et systèmes externes
- Shuffle and Streaming Engine

Objectifs

- 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.

Activités

- Quiz

Module 18: Testing et CI/CD

Sujets

- Présentation des tests et CI/CD
- Tests unitaires
- Tests d'intégration
- Construction d'artefacts
- Déploiement

Objectifs

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

Activités

- Lab pratique et quiz

Module 19: Fiabilité

Sujets

- Introduction à la fiabilité
- Surveillance
- Géolocalisation
- Reprise après sinistre
- Haute disponibilité

Objectifs

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

Activités

- Quiz

Module 20: Flex Templates

Sujets

- Modèles classiques
- Modèles flexibles
- Utiliser les Flex Templates
- Modèles fournis par Google

Objectifs

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

Activités

- Lab pratique et quiz

Module 21: Conclusion

Sujets

- Synthèse

Objectifs

- Récapitulatif rapide des sujets de formation