Formation Serverless Data Processing with Dataflow

Approfondissez votre maîtrise de Dataflow

N° ACTIVITÉ : 11 92 18558 92 | TÉLÉPHONE : 01 85 77 07 07 | E-MAIL : inscription@hubformation.com

Cette formation est destinée aux praticiens du Big Data qui souhaitent approfondir leur compréhension de Dataflow afin de faire progresser leurs applications de traitement de données.
En commençant par les bases, cette formation explique 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.
La section sur le développement de pipelines explique comment convertir votre logique métier en applications de traitement de données pouvant s'exécuter sur Dataflow.
Cette 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.

Référence	GCP300DATAFLOW
Durée	3 jours (21h)
Tarif	2 100 €HT

SESSIONS PROGRAMMÉES

A DISTANCE (FRA)

du 21 au 23 juillet 2025
du 6 au 8 octobre 2025

VOIR TOUTES LES DATES

Objectifs

| 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 oeuvre 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.

Prérequis

| Avoir suivi la formation " Google Cloud Platform - Ingénierie de données " ou avoir des connaissances équivalentes

Public

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

Programme de la formation

Introduction
| Actualisation des faisceaux et des flux de données
| Démontrer comment Apache Beam et Dataflow fonctionnent ensemble pour répondre aux besoins de traitement des données de votre organisation.

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

Séparer le calcul et le stockage avec Dataflow
| Dataflow
| Dataflow Shuffle Service
| Dataflow Streaming Engine
| Flexible Resource Scheduling
| 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.

IAM, Quotas et Permissions
| IAM
| Quota
| 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.

Sécurité
| Localité des données
| Shared VPC
| IPs privées
| CMEK
| 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 oeuvre les meilleures pratiques pour un environnement de traitement de données sécurisé.

Revue des concepts de BEAM
| Les bases Beam
| Transformations utilitaires
| Cycle de vie DoFn
| 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.

Windows, Watermarks, Triggers
| 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).

Sources and Sinks
| Sources et Sinks
| Text IO et File IO
| BigQuery IO
| PubSub IO
| Kafka IO
| Bigable IO
| Avro IO
| Splittable DoFn
| 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.

Schémas
| Beam Schemas
| Exemples de code
| Introduire des schémas, qui donnent aux développeurs un moyen d'exprimer des données structurées dans leurs pipelines Beam.

HUB Formation - 14 avenue de l'Opéra - 75001 Paris
01 85 77 07 07 - www.hubformation.com - SIREN 525 287 785

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

État et Timers
| State API
| Timer API
| Summary
| 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

Bonnes pratiques
| 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
| Implement best practices for Dataflow pipelines.

Dataflow SQL et DataFrames
| Dataflow et Beam SQL
| Windowing in SQL
| Beam DataFrames
| Développer un pipeline Beam en utilisant SQL et DataFrames.

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

Monitoring
| Job List
| Job Info
| Job Graph
| Job Metrics
| Metrics Explorer

Monitoring
| Logging
| Rapport d'erreur
| Utiliser les journaux Dataflow et les widgets de diagnostic pour résoudre les problèmes de pipeline.

Dépannage et débogage
| Flux de travail de dépannage
| Types de problèmes
| Utiliser une approche structurée pour déboguer vos pipelines Dataflow.
| Examiner les causes courantes des défaillances de pipeline.

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

Testing et CI/CD
| Présentation des tests et CI/CD
| Tests unitaires
| Tests d'intégration
| Construction d'artefacts
| Déploiement
| 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.

HUB Formation - 14 avenue de l'Opéra - 75001 Paris
01 85 77 07 07 - www.hubformation.com - SIREN 525 287 785

Fiabilité
| Introduction à la fiabilité
| Surveillance
| Géolocalisation
| Reprise après sinistre
| Haute disponibilité
| Mettre en oeuvre les bonnes pratiques en matière de fiabilité pour vos pipelines Dataflow.

Flex Templates
| Modèles classiques
| Modèles flexibles
| Utiliser les Flex Templates
| Modèles fournis par Google
| Utiliser des Flex Templates pour standardiser et réutiliser le code du pipeline Dataflow.

Méthode pédagogique

Chaque participant travaille sur un poste informatique qui lui est dédié. Un support de cours lui est remis soit en début soit en fin de cours. La théorie est complétée par des cas pratiques ou exercices corrigés et discutés avec le formateur. Le formateur projette une présentation pour animer la formation et reste disponible pour répondre à toutes les questions.

Méthode d'évaluation

Tout au long de la formation, les exercices et mises en situation permettent de valider et contrôler les acquis du stagiaire. En fin de formation, le stagiaire complète un QCM d'auto-évaluation.

Suivre cette formation à distance

Voici les prérequis techniques pour pouvoir suivre le cours à distance :

| Un ordinateur avec webcam, micro, haut-parleur et un navigateur (de préférence Chrome ou Firefox). Un casque n'est pas nécessaire suivant l'environnement.
| Une connexion Internet de type ADSL ou supérieure. Attention, une connexion Internet ne permettant pas, par exemple, de recevoir la télévision par Internet, ne sera pas suffisante, cela engendrera des déconnexions intempestives du stagiaire et dérangera toute la classe.
| Privilégier une connexion filaire plutôt que le Wifi.
| Avoir accès au poste depuis lequel vous suivrez le cours à distance au moins 2 jours avant la formation pour effectuer les tests de connexion préalables.
| Votre numéro de téléphone portable (pour l'envoi du mot de passe d'accès aux supports de cours et pour une messagerie instantanée autre que celle intégrée à la classe virtuelle).
| Selon la formation, une configuration spécifique de votre machine peut être attendue, merci de nous contacter.
| Pour les formations incluant le passage d'une certification la dernière journée, un voucher vous est fourni pour passer l'examen en ligne.
| Pour les formations logiciel (Adobe, Microsoft Office...), il est nécessaire d'avoir le logiciel installé sur votre machine, nous ne fournissons pas de licence ou de version test.
| Horaires identiques au présentiel.

Accessibilité

Les sessions de formation se déroulent sur des sites différents selon les villes ou les dates, merci de nous contacter pour vérifier l'accessibilité aux personnes à mobilité réduite.
Pour tout besoin spécifique (vue, audition...), veuillez nous contacter au 01 85 77 07 07.