Concevoir et Optimiser des Pipelines Serverless avec Apache Beam et Dataflow

Lien : https://innov-systems.com/formation/concevoir-et-optimiser-des-pipelines-serverless-avec-apache-beam-et-dataflow

DURÉE
4 jours (28h)

RÉFÉRENCE
VSC309

CATÉGORIE
Google Cloud

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre les fondements d’Apache Beam et de Google Cloud Dataflow pour le traitement serverless des données
- Concevoir, développer et exécuter des pipelines performants et sécurisés en mode batch et streaming
- Maîtriser la gestion des ressources, la portabilité, la sécurité et la performance dans un environnement Dataflow
- Mettre en œuvre des pipelines Dataflow intégrés à une démarche CI/CD et supervisés par les outils de monitoring Cloud
- Acquérir les bonnes pratiques d’architecture, de fiabilité et d’optimisation pour des projets de data engineering en production

POUR QUI ?

- Data engineers
- Architectes Cloud
- Développeurs orientés données
- Responsables techniques de plateformes data

Programme détaillé

1 / INTRODUCTION À L'APPROCHE SERVERLESS ET À DATAFLOW

- Comprendre le paradigme serverless appliqué au traitement de données massives
- Identifier les cas d'usage typiques de Google Cloud Dataflow
- Comparer les approches traditionnelles de data processing avec Beam/Dataflow

2 / ARCHITECTURE D'APACHE BEAM

- Découvrir la structure d'un pipeline Beam : PCollection, PTransform, Runner
- Comprendre le cycle de vie d'un pipeline et le rôle des DoFn
- Identifier les principaux SDK disponibles (Python, Java, Go)

3 / PREMIER PIPELINE DATAFLOW

- Créer un pipeline Beam simple et l'exécuter sur Dataflow
- Configurer les paramètres d'exécution et le choix du runner
- Comprendre la journalisation et le suivi des jobs dans la console GCP

4 / CONCEPTS CLÉS : WINDOWS, WATERMARKS ET TRIGGERS

- Implémenter des fenêtres temporelles pour le traitement en continu
- Gérer les données tardives et les déclencheurs d'événements
- Utiliser les bonnes stratégies de traitement pour unbounded data

5 / OPTIMISATION DES PIPELINES

- Identifier les goulots d'étranglement de performance
- Comprendre l'impact des partitions, du shuffle et du streaming engine
- Utiliser le planificateur flexible pour ajuster les coûts et les performances

Innov Systems

6 / GESTION DU STOCKAGE ET DU CALCUL

- Séparer le calcul du stockage avec Shuffle et Streaming Engine
- Configurer les E/S : BigQuery, Pub/Sub, Cloud Storage, Bigtable
- Implémenter des sources et sinks personnalisés via Splittable DoFn (SDF)

7 / GESTION DES DROITS ET DE LA SÉCURITÉ

- Définir les rôles IAM et permissions nécessaires à l'exécution des pipelines
- Mettre en place les bonnes pratiques de sécurité et de conformité
- Gérer la localisation des données et les stratégies de traitement zonales

8 / INTRODUCTION AUX SCHÉMAS ET STRUCTURATION DES DONNÉES

- Déclarer et manipuler des schémas dans Beam
- Simplifier le code et améliorer les performances grâce aux schémas
- Comprendre la sérialisation et les implications sur les performances

9 / MONITORING ET DIAGNOSTIC

- Surveiller les exécutions de pipeline via l'interface Dataflow
- Interpréter les métriques et logs pour détecter les anomalies
- Créer des tableaux de bord et des alertes avec Cloud Monitoring

10 / DÉBOGAGE ET DÉPANNAGE STRUCTURÉ

- Identifier les causes courantes d'échec de pipeline
- Utiliser les journaux et widgets de diagnostic pour corriger les erreurs
- Appliquer une méthodologie de débogage efficace pour Beam

11 / TESTS ET QUALITÉ DES PIPELINES

- Mettre en œuvre des tests unitaires et d'intégration pour Beam
- Automatiser les tests avec les frameworks Beam Testing et Pytest

Inov Systems

- Gérer les tests de performance et la validation des résultats

12 / INTÉGRATION CONTINUE ET DÉPLOIEMENT

- Mettre en place une chaîne CI/CD pour les pipelines Dataflow
- Gérer les templates de déploiement (Classic et Flex Templates)
- Définir des workflows de promotion entre environnements (dev, staging, prod)

13 / UTILISATION AVANCÉE DE BEAM SQL ET DATAFRAMES

- Construire des pipelines avec Beam SQL pour accélérer le prototypage
- Manipuler les DataFrames Beam pour les traitements analytiques
- Combiner SQL, Python et Java dans des pipelines hybrides

14 / NOTEBOOKS BEAM ET PROTOTYPAGE INTERACTIF

- Utiliser les notebooks Beam pour tester rapidement les transformations
- Lancer des tâches Dataflow directement depuis un environnement notebook
- Collaborer entre data scientists et data engineers sur un même pipeline

15 / FIABILITÉ ET RÉSILIENCE

- Implémenter des mécanismes de reprise et de tolérance aux pannes
- Utiliser l'état et les timers pour gérer des événements complexes
- Garantir la cohérence des traitements dans les pipelines streaming

16 / BONNES PRATIQUES ET GOUVERNANCE DATAFLOW

- Standardiser les déploiements avec les Flex Templates
- Structurer la documentation et les conventions de code
- Appliquer les bonnes pratiques de monitoring, sécurité et performance à l'échelle

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

17 au 20 Nov. 2026	Distanciel
05 au 08 Jan. 2027	Présentiel - Casablanca
02 au 05 Mars 2027	Présentiel - Casablanca
27 au 30 Avr. 2027	Présentiel - Casablanca
22 au 25 Juin 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés