Formation Dataflow & Streaming Analytics

Informations

Durée : 3 jours (21h.)

Tarif* : Nous consulter

Réf : GCSA

Niveau : Moyen

intra

Mise à jour le 29/09/25

*tarif valable jusqu'au 31/12/2025

Prochaines sessions

Contactez-nous pour connaitre nos futures sessions.

Pré-requis

- Connaissances de base en SQL et bases de données
- Notions de Python ou Java (Apache Beam
- Familiarité avec les services GCP (BigQuery, Cloud Storage)

Objectifs

Objectifs pédagogiques :

- Comprendre les concepts du traitement batch et streaming dans GCP
- Découvrir l'écosystème de traitement de données temps réel : Pub/Sub, Dataflow, BigQuery
- Construire et gérer des pipelines de données avec Apache Beam et Dataflow
- Superviser, sécuriser et optimiser des flux de données en production

Objectifs opérationnels :

- Concevoir, déployer et exploiter un pipeline de données en temps réel sur GCP : ingestion via Pub/Sub, traitement avec Dataflow et Apache Beam, stockage dans BigQuery, et supervision / optimisation des performances et des coûts.

Programme

Jour 1 – Fondamentaux du Streaming & Dataflow

Introduction au traitement de données : batch vs streaming, différences, cas d'usage
Présentation de l'écosystème Data GCP : Pub/Sub, Dataflow, BigQuery, Dataproc
Apache Beam Fundamentals : PCollections, PTransforms, Pipelines
Sources & Sinks (Cloud Storage, Pub/Sub, BigQuery)
Fenêtrage et parallélisme
Introduction à Dataflow : architecture serverless et gestion des workers
Différences entre Dataflow et Spark/Hadoop
Projet pratique : créer un pipeline Dataflow batch (CSV depuis Cloud Storage vers BigQuery). Vérification des coûts et logs

Jour 2 – Streaming temps réel & Intégrations avancées

Ingestion en temps réel avec Pub/Sub : publishers, subscribers, topics, subscriptions
Intégration avec Dataflow
Pipelines Dataflow pour le streaming : fenêtrage (fixed, sliding, session), triggers, gestion du retard (lateness), watermarks
Enrichissement et transformation des données : jointures stream-stream, stream-batch, nettoyage et normalisation
Intégrations multi-services : BigQuery (analytique temps réel), Cloud Storage (archivage), Looker Studio (visualisation)
Projet pratique : pipeline Pub/Sub → Dataflow → BigQuery pour ingestion temps réel de logs applicatifs

Jour 3 – Supervision, Optimisation & Cas pratiques

Supervision & Debugging : utiliser la console Dataflow, logs et métriques (latence, throughput), Stackdriver Logging
Optimisation des performances et coûts : autoscaling horizontal, parallélisme, bonnes pratiques pour limiter les coûts
Sécurité et IAM : rôles et permissions nécessaires pour Dataflow et Pub/Sub, gestion des secrets et accès sécurisés

Formation Dataflow & Streaming Analytics

Monitoring avec Cloud Logging et Cloud Monitoring
Projet fil rouge : mise en place d’un pipeline temps réel complet avec ingestion Pub/Sub, traitement Dataflow et stockage BigQuery. Visualisation des données via Looker Studio. Analyse des coûts et mise en place d’alertes