[sfeir] Institute

Google Cloud | GCP200CDF

# Data Integration with Cloud Data Fusion

Maîtriser l'intégration de données Google Cloud à l'aide de Cloud Data Fusion

2jours / 14h

## Objectifs pédagogiques

- Identifier le besoin d'intégration de données
- Comprendre les fonctionnalités fournies par Cloud Data Fusion en tant que plate-forme d'intégration de données
- Identifier les cas d'utilisation pour une éventuelle mise en œuvre avec Cloud Data Fusion
- Répertorier les composants principaux de Cloud Data Fusion
- Concevoir et exécuter des pipelines de traitement de données par lots et en temps réel
- Travailler avec Wrangler pour créer des transformations de données
- Utiliser des connecteurs pour intégrer des données de différentes sources et formats
- Configurer l'environnement d'exécution ; Surveiller et dépanner l'exécution du pipeline
- Comprendre la relation entre les métadonnées et la lignée des données

## Public cible

- Data Engineer
- Data Analysts

# Prérequis

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

- Avoir suivi le cours « Big Data and Machine Learning Fundamentals » ou avoir des connaissances équivalentes

# Programme

**Module 0: Introduction**

Sujets

- Introduction

Objectifs

- Introduire les objectifs du cours

**Module 1: Introduction à l'intégration de données et Cloud Data Fusion**

Sujets

- Intégration de données : quoi, pourquoi, défis
- Outils d'intégration de données utilisés dans l'industrie
- Personas utilisateur
- Introduction à la fusion de données cloud
- Capacités critiques d'intégration de données
- Composants de l'interface utilisateur Cloud Data Fusion

Objectifs

- Comprendre le besoin d'intégration de données
- Lister les situations/cas où l'intégration de données peut aider les entreprises
- Lister les plateformes et outils d'intégration de données disponibles
- Identifier les défis liés à l'intégration des données
- Comprendre l'utilisation de Cloud Data Fusion en tant que plate-forme d'intégration de données
- Créer une instance Cloud Data Fusion
- Se familiariser avec le framework de base et les principaux composants de Cloud Data Fusion

Activités

- Lab noté, quiz, discussions

**Module 2: Construire des pipelines**

Sujets

- Architecture de Cloud Data Fusion
- Concepts de base
- Pipelines de données et graphes acycliques dirigés (DAG)
- Cycle de vie des pipelines
- Conception de pipelines dans Pipeline Studio

Objectifs

- Comprendre l'architecture de Cloud Data Fusion
- Définir ce qu'est un pipeline de données
- Comprendre la représentation DAG d'un pipeline de données,
- Apprendre à utiliser Pipeline Studio et ses composants
- Concevoir un pipeline simple à l'aide de Pipeline Studio,
- Déployer et exécuter un pipeline

Activités

- Lab noté, quiz

**Module 3: Construire des pipelines complexes**

Sujets

- Branchement, fusion et jointure
- Actions et notifications
- Gestion des erreurs et macros
- Configurations de pipeline, planification, importation et exportation

Objectifs

- Effectuer des opérations de branchement, de fusion et de jointure.
- Exécuter le pipeline avec des arguments d'exécution à l'aide de macros.
- Travailler avec des gestionnaires d'erreurs.
- Exécuter des exécutions pré- et post-pipeline à l'aide d'actions et de notifications.
- Planifier l'exécution des pipelines.
- Importer et exporter des pipelines existants.

Activités

- Lab noté, quiz

**Module 4: Environnement d’exécution du pipeline**

Sujets

- Horaires et déclencheurs
- Environnement d’exécution : profil de calcul et provisionneurs
- Surveillance des pipelines

Objectifs

- Comprendre la composition d’un environnement d’exécution.
- Configurer l’environnement d’exécution, la journalisation et les métriques de votre pipeline. Comprendre des concepts tels que le profil de calcul et l’approvisionneur.
- Créer un profil de calcul.
- Créer des alertes de pipeline.
- Surveiller le pipeline en cours d’exécution.

Activités

- Quiz

**Module 5: Construire des transformations et préparer des données avec Wrangler**

Sujets

- Wrangler
- Directives
- Directives définies par l’utilisateur

Objectifs

- Comprendre l’utilisation de Wrangler et de ses principaux composants.
- Transformer les données à l’aide de l’interface utilisateur Wrangler.
- Transformer les données à l’aide de directives/méthodes CLI.
- Créer et utiliser des directives définies par l’utilisateur.

Activités

- Lab noté, quiz

**Module 6: Connecteurs et pipelines de streaming**

Sujets

- Comprendre l’architecture d’intégration de données.
- Lister les différents connecteurs.
- Utilisez l’API Cloud Data Loss Prevention (DLP).
- Comprendre l’architecture de référence des pipelines de streaming.
- Construire et exécuter un pipeline de streaming.

Objectifs

- Connecteurs
- DLP
- Architecture de référence pour les applications de streaming
- Création de pipelines de diffusion en continu

Activités

- Lab noté, quiz, discussions

**Module 7: ??Métadonnées et lignage des données**

Sujets

- Métadonnées
- Lignage des données

Objectifs

- Répertorier les types de métadonnées.
- Différencier les métadonnées commerciales, techniques et opérationnelles.
- Comprendre ce qu’est le lignage des données.
- Comprendre l’importance de maintenir la lignée des données.
- Différencier les métadonnées et le lignage des données.

Activités

- Lab noté, quiz

**Module 9: Conclusion**

Sujets

- Synthèse

Objectifs

- Revue des objectifs et des concepts du cours