

## Data Engineering on AWS

**Durée: 3 Jours    Réf de cours: GK910032    Méthodes d'apprentissage: Classe à distance**

### Résumé:

**Data Engineering on AWS is a 3-day intermediate course, designed for professionals seeking a deep dive into data engineering practices and solutions on AWS.**

Through a balanced combination of theory, practical labs, and activities, participants learn to design, build, optimize, and secure data engineering solutions using AWS services. From foundational concepts to hands-on implementation of data lakes, data warehouses, and both batch and streaming data pipelines, this course equips data professionals with the skills needed to architect and manage modern data solutions at scale.

- **Course level:** Intermediate

- **Duration:** 3 day

*Updated August 2026*

Classe à Distance - site Client

Cette formation peut être suivie à distance en synchrone depuis n'importe quel site pourvu d'une connexion internet (2 Mb/s en symétrique recommandés). Le programme (théorie et pratique) suit le même déroulé pédagogique qu'en présentiel. La solution technologique adoptée permet aux apprenants à distance de suivre les présentations faites au tableau, de voir et d'entendre l'instructeur et les participants en temps réel, mais également d'échanger avec eux.

### Public visé:

This course is designed for professionals who are interested in designing, building, optimizing, and securing data engineering solutions using AWS services.

### Objectifs pédagogiques:

#### ■ In this course, you will learn to do the following:

- Understand the foundational roles and key concepts of data engineering, including data personas, data discovery, and relevant AWS services.
- Identify and explain the various AWS tools and services crucial for data engineering, encompassing orchestration, security, monitoring, CI/CD, IaC, networking, and cost optimization.
- Design and implement a data lake solution on AWS, including storage, data ingestion, transformation, and serving data for consumption.
- Optimize and secure a data lake solution by implementing open table formats, security measures, and troubleshooting common issues.
- Design and set up a data warehouse using Amazon Redshift Serverless, understanding its architecture, data ingestion, processing, and serving capabilities.
- Apply performance optimization techniques to data warehouses in Amazon Redshift, including monitoring, data optimization, query optimization, and orchestration.
- Manage security and access control for data warehouses in Amazon Redshift, understanding authentication, data security, auditing, and compliance.
- Design effective batch data pipelines using appropriate AWS services for processing and transforming data.
- Implement comprehensive strategies for batch data pipelines, covering data processing, transformation, integration, cataloging, and serving data for consumption.
- Optimize, orchestrate, and secure batch data pipelines, demonstrating advanced skills in data processing automation and security.
- Architect streaming data pipelines, understanding various use cases, ingestion, storage, processing, and analysis using AWS services.
- Optimize and secure streaming data solutions, including compliance considerations and access control.

## Pré-requis:

### We recommend that attendees of this course have:

- Familiarity with basic machine learning concepts, such as supervised and unsupervised learning, regression, classification, and clustering algorithms.
- Working knowledge of Python programming language and common data science libraries like NumPy, Pandas, and Scikit-learn.
- Basic understanding of cloud computing concepts and familiarity with the AWS platform.
- Familiarity with SQL and relational databases is recommended but not mandatory.
- Experience with version control systems like Git is beneficial but not required.

## Test et certification

None

---

### Après cette formation, nous vous conseillons le(s) module(s) suivant(s):

- GK0630 - Mise en pratique de la Data Science avec Amazon SageMaker
- GK910029 - Machine Learning Engineering on AWS

## Contenu:

### Day 1

#### Module 1: Data Engineering Roles and Key Concepts

- Role of a Data Engineer
- Key functions of a Data Engineer
- Data Personas
- Data Discovery
- AWS Data Services

#### Module 2: AWS Data Engineering Tools and Services

- Orchestration and Automation
- Data Engineering Security
- Monitoring
- Continuous Integration and Continuous Delivery
- Infrastructure as Code
- AWS Serverless Application Model
- Networking Considerations
- Cost Optimization Tools

#### Module 3: Designing and Implementing Data Lakes

- Data lake introduction
- Data lake storage
- Ingest data into a data lake
- Catalog data
- Transform data
- Server data for consumption
- Hands-on lab: Setting up a Data Lake on AWS

#### Module 4: Optimizing and Securing a Data Lake Solution

- Open Table Formats
- Security using AWS Lake Formation
- Setting permissions with Lake Formation
- Security and governance
- Troubleshooting
- Hand-on lab: Automating Data Lake Creation using AWS Lake Formation Blueprints

### Day 2

#### Module 5: Data Warehouse Architecture and Design Principles

- Introduction to data warehouses
- Amazon Redshift Overview
- Ingesting data into Redshift
- Processing data
- Serving data for consumption
- Hands-on Lab: Setting up a Data Warehouse using Amazon Redshift Serverless

#### Module 6: Performance Optimization Techniques for Data Warehouses

- Monitoring and optimization options
- Data optimization in Amazon Redshift
- Query optimization in Amazon Redshift
- Orchestration options

#### Module 7: Security and Access Control for Data Warehouses

- Authentication and access control in Amazon Redshift
- Data security in Amazon Redshift
- Auditing and compliance in Amazon Redshift
- Hands-on lab: Managing Access Control in Redshift

#### Module 8: Designing Batch Data Pipelines

- Introduction to batch data pipelines
- Designing a batch data pipeline
- AWS services for batch data processing

### Module 9: Implementing Strategies for Batch Data Pipeline

- Elements of a batch data pipeline
- Processing and transforming data
- Integrating and cataloging your data
- Serving data for consumption
- Hands-on lab: A Day in the Life of a Data Engineer

### Day 3

#### Module 10: Optimizing, Orchestrating, and Securing Batch Data Pipelines

- Optimizing the batch data pipeline
- Orchestrating the batch data pipeline
- Securing the batch data pipeline
- Hands-on lab: Orchestrating Data Processing in Spark using AWS Step Functions

#### Module 11: Streaming Data Architecture Patterns

- Introduction to streaming data pipelines
- Ingesting data from stream sources
- Streaming data ingestion services
- Storing streaming data
- Processing Streaming Data
- Analyzing Streaming Data with AWS Services
- Hands-on lab: Streaming Analytics with Amazon Managed Service for Apache Flink

#### Module 12: Optimizing and Securing Streaming Solutions

- Optimizing a streaming data solution
- Securing a streaming data pipeline
- Compliance considerations
- Hands-on lab: Access Control with Amazon Managed Streaming for Apache Kafka

## Méthodes pédagogiques :

### Activities:

This course includes presentations, demonstrations, hands-on labs, and group exercises.

## Autres moyens pédagogiques et de suivi:

- **Compétence du formateur :** Les experts qui animent la formation sont des spécialistes des matières abordées et ont au minimum cinq ans d'expérience d'animation. Nos équipes ont validé à la fois leurs connaissances techniques (certifications le cas échéant) ainsi que leur compétence pédagogique.
- **Suivi d'exécution :** Une feuille d'émargement par demi-journée de présence est signée par tous les participants et le formateur.
- **En fin de formation,** le participant est invité à s'auto-évaluer sur l'atteinte des objectifs énoncés, et à répondre à un questionnaire de satisfaction qui sera ensuite étudié par nos équipes pédagogiques en vue de maintenir et d'améliorer la qualité de nos prestations.