DataTips

Formation

AWS Data Engineering

DataTipsLearning.com

Tarif :
2000 € HT

Durée :
4 Jours
(28h)

Lieu :
PARIS/
Distancielle/
Votre Local

Contact :

contact@datatipslearning.com
+33 0751386021
Réf. DTPSL-AWSDG-2022
S'inscrire

Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

AWS est une plateforme cloud développée par Amazon. Il regroupe plus de 100 services répartis en diverses catégories telles que le stockage, le calcul, l'analyse de données, et l'intelligence artificielle.

Lancé en 2006, AWS est Initialement conçu comme une ressource interne, est ensuite devenu un fournisseur de solutions cloud innovantes et économiques permettant aux entreprises qui les utilisent de grandir et de monter en échelle.

Objectifs

- Présenter les services AWS pour le Data Engineering.
- Apprendre à analyser les données avec les services AWS.
- Apprendre le calcul avec Elastic Compute Cloud EC2.
- Transformer les données avec un cluster Big Data du service EMR.
- Apprendre à utiliser un Data Warehouse avec AWS Redshift.
- Comprendre à orchestrer un pipeline de données avec AWS pipeline & AWS Step Function.
- Requêter les données avec Amazon Athena.
- Travailler avec les différents types de service de stockage de AWS comme S3, EBS et EFS.
- Utiliser des bases de données AWS comme RDS et Document DB.
- Ingestion des données en batch et en streaming.
- Apprendre à gouverner, sécuriser et cataloguer les données.

Public

- Data Engineer
- Data Architectes
- Data Scientist
- Data Analyst

Prérequis

- Connaissances des bases de AWS.
- Quelques connaissances de SQL.

Programme détaillé

I. Introduction au Data Engineering
   1. Présentation du Big Data
   2. Les avantages du cloud pour des solutions d'analyse big data

II. Architectures du data management pour l'analytique
   1. L'évolution du data management pour l'analytique
   2. Comprendre data Warehouse & data Mart
   3. Construire un data Lake pour maîtriser la variété et le volume du big data

4. L'architecture lakehouse

III. Commencer à utiliser AWS
   1. Présentation de Amazon Web Services
   2. S'inscrire à Amazon Web Services
   3. Créer un utilisateur et des groupes
   4. Créer des paires de clés
   5. Installer et configurer AWS CLI
   6. Télécharger le client SSH
   7. Configurer Éclipse
   8. Création et accès à un compte AWS
   9. Création d'un bucket S3

IV. Les services AWS pour le Data Engineering
   1. Services AWS pour l'ingestion de données
   2. Services AWS pour consommer des données
   3. Services AWS pour transformer les données
   4. Services AWS pour orchestrer les pipelines Big Data
   5. Déclenchement d'une fonction AWS Lambda quand un fichier arrive dans un bucket S3

V. Architecture des pipelines de Data Engineering
   1. L'architecture data pipeline
   2. Identifier les consommateurs de données
   3. Identifier les sources de données
   4. Identifier les transformations et optimisations des données
   5. Charger des données dans des data Marts
   6. Architecture d'un exemple de pipeline

VI. Ingestion des données en batch et en streaming
   1. Comprendre les sources de données
   2. Ingestion de données à partir d'une base de données relationnelle
   3. Ingestion des données de streaming
   4. Ingestion des données avec AWS DMS
   5. Transférer données de S3 à S3 avec Amazon AppFlow

VII. Transformer les données pour optimiser l'analyse
   1. Extraire les valeurs à partir des données brutes
   2. Types de service de transformation de données
   3. Transformations pour la préparation des données
   4. Exemple métier de transformations
   5. Utilisation change data capture (CDC)
   6. Joindre des ensembles de données avec AWS Glue Studio
   7. Traitement de données S3 avec AWS Batch
   8. Transformation ETL avec AWS Lambda

VIII. EMR pour effectuer des analytiques Big Data

1. Introduction à EMR
2. Création d'un cluster EMR
3. Soumission d'un job à un cluster EMR
4. Traitement des données avec Spark
5. Exécution des étapes EMR

IX. Identification et activation des consommateurs de données
1. Comprendre la démocratisation des données
2. Répondre aux besoins métiers avec la visualisation des données
3. Répondre aux besoins des analystes avec un reporting structuré
4. Répondre aux besoins des data scientists et des modèles ML
5. Création de transformations de données avec AWS Glue DataBrew

X. Chargement de données dans un data Mart
1. L'analyse avec les entrepôts de données/data Marts
2. Les anti-modèles pour un entrepôt de données
3. L'architecture Redshift et analyse approfondie du stockage
4. Conception d'un data Warehouse performant
5. Déplacer des données un data Lake et Redshift
6. Chargement de données dans un cluster Amazon Redshift et exécution de requêtes

XI. Orchestration du pipeline de données
1. Les concepts de base de l'orchestration de pipeline
2. Les options d'orchestration des pipelines dans AWS
3. Orchestrer un pipeline de données avec AWS Step Function
4. Automatiser le mouvement et la transformation des données avec AWS Data Pipeline

XII. Requêtes ad hoc avec Amazon Athena
1. Analyse SQL pour Data Lake avec Amazon Athena
2. Optimiser les requêtes Amazon Athena
3. Fédérer les requêtes de sources de données externes avec Amazon Athena Query Federation
4. Gestion de la gouvernance et des coûts avec Amazon Athena workgroups
5. Création et configuration de Amazon Athena workgroup
6. Changer workgroups et exécuter des requêtes

XIII. Elastic Compute Cloud EC2
1. Introduction à EC2
2. Fonctions et caractéristiques de EC2
3. Types d'instance EC2
4. Gestion d'EC2 et des instances
5. La paire de clés et les groupe de sécurité
6. Monitoring avec CloudWatch

XIV. Visualisation des données avec Amazon QuickSight
1. Représentation graphique des données
2. Les concepts de base d'Amazon QuickSight
3. Ingestion et préparation des données de diverses sources
4. Création et partage des dashboards QuickSight

XV. Traitement de données en temps réel avec Amazon Kinesis
1. Présentation de Kinesis streaming data platform
2. Présentation de Kinesis Stream et Firehose
3. Envoie de données avec Kinesis Data Firehose
4. Présentation de Kinesis Data Analytics
5. Les concepts de Streaming SQL
6. Charger, traiter et stocker les données en temps quasi réel à l'aide de requêtes SQL standard
7. Traiter des données Kinesis avec Apache Spark

XVI. Stockage AWS
1. Présentation de S3
2. Concepts et utilisations de S3
3. Concepts et utilisations de EBS
4. Utilisation des politiques de cycle de vie S3 pour réduire les coûts de stockage
5. Archiver automatiquement les objets S3 avec les stratégies d'archivage
6. Réplication des buckets S3 pour la récupération
7. Les métriques de stockage et d'accès S3 à l'aide de Storage Lens
8. Configuration de l'accès aux buckets S3
9. Chiffrer des objets avec KMS
10. Création et restauration de sauvegardes EC2 dans une autre région avec AWS Backup
11. Restauration d'un fichier d'un Snapshot EBS
12. Utilisation des Instance Store de EC2
13. Réplication de données entre EFS et S3 avec DataSync
14. Utiliser AWS Elastic Disaster Recovery

XVII. Les bases de données AWS
1. Présentation des bases de données AWS
2. Création d'une base de données PostgreSQL serverless Aurora
3. Utilisation de l'authentification IAM avec une base de données RDS
4. Utiliser proxy RDS pour les connexions de base de données à partir de Lambda
5. Chiffrement du stockage d'une base de données RDS MySQL
6. Automatisation de la rotation des mots de passe pour les bases de données RDS
7. Autoscaling automatique d'une table Dynamo DB
8. Migration de bases de données vers Amazon RDS via AWS DMS

9. Activation de l'accès REST à Aurora serverless avec RDS Data API
10. Créer un cluster Document DB et requêter des données

XVIII. Catalogue, sécurité et gouvernance
1. Assurer la sécurité et la gouvernance des données
2. Cataloguer les données pour éviter data swamp
3. Le catalogue de données AWS Glue/Lake Formation
4. Les services AWS pour le chiffrement des données et le monitoring de sécurité
5. Les services AWS pour la gestion de l'identité et des autorisations
6. Configuration des autorisations de Lake Formation

Méthode pédagogique

La formation se compose d'une partie théorique, et également une partie pratique représentant 60% de de la formation.
La partie pratique contient plusieurs exercices sous forme de notebook Databricks avec les corrections, avec aussi un projet à la fin de la formation comme simulation d'une prod.

Chaque jour, une évaluation rapide des connaissances est effectuée avant de commencer les nouvelles parties de la formation.
A la fin, une synthèse globale est délivré aux stagiaires, renforcé par un projet prod.
Finalement, une évaluation QCM est proposée.

Un support de cours sera remis à chaque stagiaire comprenant les slides, les exercices et les corrigés et un git du projet prod.
Une feuille de présence est fournie en fin de formation avec une certificat de complétion de formation pour chaque stagiaire.

Le formateur est un Data Engineer expert, qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil.