alter way

FORMATION POUR DATA INGÉNIEUR CLOUD AWS

Durée 3 jours (21h)

Cette formation s'adresse aux Data Ingénieurs confirmés ayant déjà de l'expérience dans un environnement data On-Premise (MapR, Cloudera ou Hortonworks), ou ayant une première expérience avec AWS et souhaitant l'approfondir.

Public concerné :

- Data Ingénieurs Cloud Confirmés
- Cloud Ingénieurs avec une appétence développement

Prérequis nécessaires :

- Base système / Linux
- Avoir déjà utilisé un langage de programmation auparavant (Java, Python, C++, ...)
- Connaître les bases du cloud et des services « serverless »

- **Objectifs de la formation :**
- Apprendre à mettre en place un pipeline de données de bout en bout sur le cloud AWS.
- Avoir en sortie de la formation un projet concret, mis en place de bout en bout.
- Connaître l'environnement Cloud du Data Ingénieur

Matériel pédagogique

<u>En présentiel :</u>

Les formations sont dispensées dans des salles de formation équipés d'ordinateurs portables par défaut sous Linux (Ubuntu), si besoin sur Windows. Les participants ont accès à internet en wifi ou via des câbles Ethernet.

Le support de cours est projeté dans la salle de formation via un vidéoprojecteur, remis au stagiaire s'il apporte une clé USB, ou encore envoyé par email après la formation (sur demande). Le formateur dispose d'un paperboard pour détailler ou insister sur certains aspects. Un bloc-notes et un stylo sont mis à disposition du participant.

<u>En distanciel :</u>

Les formations sont dispensées à distance via un outil LMS et en langue française.

Le support de cours est partagé à tous les participants via l'outil de visioconférence (et/ou plateforme LMS) et envoyé par e-mail après la formation (sur demande).

Le formateur est en interaction directe avec les stagiaires via l'outil de LMS et l'outil de visioconférence utilisé.

Pédagogie

Les cours théoriques seront dispensés en alternance avec des cas pratiques afin de confronter le participant à diverses situations et lui apprendre à acquérir les bons réflexes et les bonnes pratiques.

Moyens d'encadrement / Suivi de l'exécution de l'action

- Le programme de la formation est remis aux participants avant leur inscription
- Une attestation de formation est établie et transmise au participant quelques jours après la formation.

Évaluation

Chaque participant est évalué au cours de la formation au travers des différents travaux pratiques proposés, appelés « LABS ». Un questionnaire de satisfaction est complété par les participants (avec et sans le formateur afin de leur laisser la possibilité d'exprimer librement leurs remarques) en fin de formation.

Cette évaluation est ensuite adressée au commercial en charge du client afin qu'il en prenne connaissance et puisse mesurer la satisfaction client.

alter way

PROGRAMME DES 3 JOURS –
FORMATION POUR DATA INGÉNIEUR CLOUD AWS

Fondamentaux des Systèmes Distribués – Hadoop

- Introduction aux Systèmes et Calculs Distribués.
- Stockage Distribué dans Hadoop avec HDFS
- Calcul Distribué avec YARN et gestion des ressources.

Elastic MapReduce

- Découvrir Hadoop à travers le service EMR de AWS.
- Mise en place d'un Cluster EMR, choix des machines et pricing.
- Savoir sizer le Cluster EMR dépendamment du volume de données et du workload.
- Découverte des services accompagnant EMR.
- EMR vs Glue pour les jobs Spark.
- Lancer un Cluster EMR grâce à l'API Boto3 et Lambda.
- Limites de EMR.
- Les bonnes pratiques pour réussir son utilisation du service EMR.

Ingestion des Données – Kinesis et Kafka

- Découverte des services Kinesis Data Streams et Firehose
- Découverte du service Managé Kafka de AWS.
- MSK vs Kinesis.
- Ingestion des données en Batch.
- Ingestion des données en Streaming.

Stockage des données – Datalake dans AWS

- Utilisation du Stockage Objet – S3 et ses variantes.
- Mise en place d'un Datalake sur S3 – Les bonnes pratiques et les erreurs à éviter.
- Structuration du Datalake sur S3 – Données et Métadonnées.
- Les types de fichiers – ORC, Parquet, Avro, CSV ou JSON ?

Exposition des Données et Gestion du Data Catalog

- Découvrir AWS Athena pour analyser les données stockées sur S3.
- Configuration de Athena et gestion des droits avec IAM.
- Découvrir AWS Redshift pour analyser les données de son Datalake.
- Mise en place d'un Cluster Redshift.
- Connecter Redshift ou Athena à son outil de visualisation préféré !

Ordonnancement des Workflows

- Utilisation des Step Functions.
- Utilisation des Fonctions Lambdas.
- Aller plus loin et devenir un pro : Utilisation de Apache Airflow.
  - Mise en place de Apache Airflow sur AWS.
  - Écriture de vos premiers Dags pour l'ordonnancement de votre Data Pipeline.

Cas pratique de mise en place d'une Datapipeline de bout en bout sur AWS.