alter way

FORMATION SPARK POUR DATA

Durée 2 jours (14h)

Cette formation s'adresse aux Data Ingénieurs en sortie d'école, ceux souhaitant se reconvertir et plus globalement les débutants souhaitant avoir un premier up and running rapide autour d'Apache Spark.

Public concerné :

- Développeurs Juniors
- Développeurs Java
- Data Ingénieurs Juniors

Prérequis nécessaires :

- Base système / Linux
- Avoir déjà utilisé un langage de programmation auparavant (Java, Python, C++, ...)

Objectifs de la formation :

- Apprendre manipuler l'API Spark
- Connaître l'environnement de base du Data Ingénieur

Matériel pédagogique

En présentiel :

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.

En distanciel :

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 2 JOURS –
FORMATION SPARK POUR DATA INGÉNIEUR

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.

API Spark Low Level – Les RDD

- Découvrir le concept de distribution de calculs avec les RDD.
- Manipuler vos premiers fichiers à l'aide des RDD.
- Les opérations dans Spark : Prise en main des Transformations et des Actions dans Spark.
- L'algorithme Map Reduce appliqué à Spark.
- Les agrégations avec Spark SQL.
- Les RDD en environnement distribué.
- TP1

API Spark High Level – Spark SQL

- Connaître les différences entre les RDD et les DataFrames / DataSets.
- Différences entre les DataFrames et DataSets : pourquoi faut-il utiliser un langage typé ?
- Manipuler les DataFrames et les DataSets.
- La gestion des schémas dans Spark.
- DataFrames et DataSets dans un environnement distribué.
- Les agrégations avec Spark SQL.
- Découverte et manipulation de la SparkUI.
- TP2

Structuration des données avec Spark

- Prise en main de datasets d'exemple complexes et leur structuration.
- Utilisation des UDFs dans Spark pour les cas isolés.
- Validation des données et gestion ligne par ligne des données.
- La Data Skew dans Spark et ses remèdes.

API Streaming Spark Streaming

- Utilisation des APIs Low Level et High Level de Spark pour du Streaming.
- Micro-Batch Streaming vs Event-Based Streaming.
- Mise en place d'un flux streaming avec Apache Kafka et Spark Streaming.