Formation **Data Streaming : traitement des données en temps réel**

A l'issue de cette formation, les participants ont acquis les compétences nécessaires pour traiter des données en temps réel grâce à la maîtrise d'outils modernes comme Spark, Kafka, Airflow... 50% du temps de formation est consacré aux cas pratiques, afin de permettre aux apprenants de mettre immédiatement en application les concepts théoriques du data streaming.

Durée
4 jours

Objectifs pédagogiques

- Comprendre les spécificités du traitement de données en temps réel
- Connaître les différents composants et l'architecture d'un système de data streaming
- Construire des pipelines pour le traitement de données en continu avec Kafka, Airflow ou Spark

Public
Ingénieurs Big Data, data analysts, architectes data, data stewards...

Prérequis
Bonnes connaissances en python 3 (ou sur un autre langage de programmation orienté Back-end), un niveau intermédiaire en SQL.

Programme de formation

Introduction : principes fondamentaux du data streaming
Les avantages d'une architecture distribuée résiliente pour les systèmes de data streaming
Tolérance aux pannes, callbacks et scalabilité
Acheminement des messages entre les microservices d'un système
Suivre l'activité, les logs et collecter des mesures
Gérer des flux de données avec Kafka Streams API ou Spark Streaming
Comment les géants de la Tech utilisent le streaming dans leurs activités quotidiennes (Netflix, LinkedIn, Uber...) ?

Architecture
Gérer les sources de données (évènements, messages, logs...)
La problématique de load balacing dynamique Spark pour les pannes et la récupération
L'unification des analyses par lots (batchs), en streaming et interactives
Analytics avancée avec le Machine Learning et requêtes interactives en SQL

Cas pratiques : intégration de données en temps réel avec Databricks, Spark, Kafka ou Snowflake.

Gestion des pipelines de données Cloud avec Kafka, Airflow et Spark

Producers, consumers et concepts de réplication
Brokers, clusters, topics et partition
Le streaming de données comme moyen pour partager les données

Cas pratiques : gestion d'un data workflow avec les DAGs (Directed Acyclic Graphs) d'Airflow, gestion des brokers kafka avec Zookeeper.

Mise en œuvre d'un pipeline de données temps réel

Data streaming pour une architecture orientée évènements
Data streaming pour échantillons classiques de données
Data streaming pour les industries et l'Internet des Objets (IoT)

Projet final : construction d'un pipeline de données temps réel "from scratch" avec Kafka, Airflow, Spark, Snowflake ou Databricks (au choix des stagiaires, avec les données de leur organisation, si possible et planifié à l'avance).

Moyens et méthodes pédagogiques

- La formation alterne entre présentations des concepts théoriques et mises en application à travers d'ateliers et exercices pratiques.
- Les participants bénéficient des retours d'expérience terrains du formateur ou de la formatrice
- Un support de cours numérique est fourni aux stagiaires

Modalités d'évaluation

- **En amont de la session de formation**, un questionnaire d'auto-positionnement est remis aux participants, afin qu'ils situent leurs connaissances et compétences déjà acquises par rapport au thème de la formation.

- **En cours de formation**, l'évaluation se fait sous forme d'ateliers, exercices et travaux pratiques de validation, de retour d'observation et/ou de partage d'expérience.

- **En fin de session**, le formateur évalue les compétences et connaissances acquises par les apprenants grâce à un questionnaire reprenant les mêmes éléments que l'auto-positionnement, permettant ainsi une analyse détaillée de leur progression.