# Programme de Formation :

# BIG DATA

## Objectifs :

* Définir le concept Big Data et citer ses termes clefs.
* Décrire l'écosystème Hadoop et identifier ses différentes distributions.
* Définir les enjeux stratégiques et organisationnels des projets Big Data en entreprise.
* Utiliser la distribution Cloudera pour la découverte du Big Data.
* Connaitre l'architecture du HDFS, le fonctionnement du MapReduce.
* Distinguer les différents langages de requête de Hadoop (Hive).
* Créer et utiliser une base de données NoSQL (Hbase).
* MongoDB et ElasticSearch
* Comprendre et utiliser Apache Spark

## Programme détaillé :

Toutes les parties citées ci-dessous seront accompagnées par des workshops et/ou Ateliers.

### I. Introduction Big Data

Définir le concept Big Data
- Expliquer les caractéristiques du big data
- Citer les différents cas d'utilisation du big data
- Comment choisir et mettre en place une architecture Big Data en entreprise
- Définir HADOOP
- Présenter l'architecture de l'écosystème HADOOP

- Citer les distributions HADOOP existantes sur le marché
- Installer et Présenter la machine virtuelle Cloudera

## II. HDFS et MapReduce

Expliquer le principe du système de fichiers distribués.
- Distinguer entre Datanode et Namenode,
- Gérer des pannes,
- Gérer (insérer, modifier, supprimer) les fichiers sous HDFS.
- Comprendre le fonctionnement du MapReduce.

## III. Langages de requête Hadoop

- Distinguer les différents langages de requête de Hadoop.
- Présenter et utiliser le langage de requête HIVE.

## IV. Bases de données NoSQL

- Connaitre les caractéristiques et les avantages des BD NoSQL.
- Illustrer les différences entre une BD NoSQL et une BD relationnelle.
- Connaitre les typologies des Bases de données (clé/Valeur, Document, Graphe, Colonne).
- Créer et utiliser une base de données Hbase
- Utiliser MongoDB (collection, Document, etc)
- Découvrir ElasticSearch pour l'indexation des documents et la visualisation utilisant Kibana

## V. Apache Spark

- Introduction à Spark, Architecture et Installation
- Manipulation des RDDs( Resilient Distributed Datasets
- Manipulation des Dataframes (SparkSQL)
- Manipulation du MLlib pour les applications Machine Learning