

Big Data : architectures et infrastructures

Référence : **PYCB001**

Durée : **2 jours (14 heures)**

Certification : **Aucune**

Connaissances préalables

- Il est demandé aux participants d'avoir une bonne culture générale sur les systèmes d'information

Profil des stagiaires

- Chefs de projets, architectes, développeurs, data-scientists, et toute personne souhaitant connaître les outils et solutions pour mettre en place une architecture BigData

Objectifs

- Comprendre les concepts du BigData et savoir quelles sont les technologies implémentées
- Savoir analyser les difficultés propres à un projet BigData, les freins, les apports, tant sur les aspects techniques que sur les points liés à la gestion du projet

Certification préparée

- Aucune

Méthodes pédagogiques

- La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- Mise à disposition d'un poste de travail par participant

Formateur·rice

- Consultant-Formateur expert Bigdata

Méthodes d'évaluation des acquis

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation des compétences acquises envoyée au stagiaire
- Attestation de fin de stage adressée avec la facture

Contenu du cours

1. Introduction

- L'essentiel du BigData : calcul distribué, données non structurées.
- Besoins fonctionnels et caractéristiques techniques des projets.
- La valorisation des données.
- Le positionnement respectif des technologies de cloud, BigData et noSQL, et les liens, implications.
- Quelques éléments d'architecture.
- L'écosystème du BigData : les acteurs, les produits, état de l'art.
- Cycle de vie des projets BigData.
- Emergence de nouveaux métiers : Datascientists, Data labs, ...

2. Stockage

- Caractéristiques NoSQL : adaptabilité, extensibilité, structure de données proches des utilisateurs, développeurs
- Les types de bases de données : clé/valeur, document, colonne, graphe
- Données structurées et non structurées, documents, images, fichiers XML, JSON, CSV, ...
- Les différents modes et formats de stockage
- Stockage réparti : réplication, sharding, gossip protocol, hachage
- Systèmes de fichiers distribués : GFS, HDFS
- Quelques exemples de produits et leurs caractéristiques : Cassandra, MongoDB, CouchDB, DynamoDB, Riak, Hadoop, HBase, BigTable, ...
- Qualité des données, gouvernance de données

3. Indexation et recherche

- Moteurs de recherche
- Principe de fonctionnement
- Méthodes d'indexation
- Exemple de Lucene, et mise en œuvre avec solr
- Recherche dans les bases de volumes importants : exemples de produits et comparaison : dremel, drill, elasticsearch, MapReduce

4. Calcul et restitution, intégration

- Différentes solutions : calculs en mode batch, ou en temps réel, sur des flux de données ou des données statiques
- Les produits : langage de calculs statistiques, R Statistics Language, sas, RStudio
- Ponts entre les outils statistiques et les bases BigData
- Outils de calcul sur des volumes importants : storm en temps réel, hadoop en mode batch
- Zoom sur Hadoop : complémentarité de HDFS et MapReduce
- Restitution et analyse : logstash, kibana, elk, pentaho
- Présentation de pig pour la conception de tâches MapReduce sur une grappe Hadoop

Notre référent handicap se tient à votre disposition au [01.71.19.70.30](tel:0171197030) ou par mail à referent.handicap@edugroupe.com pour recueillir vos éventuels besoins d'aménagements, afin de vous offrir la meilleure expérience possible.