

Hadoop : analytics

Durée: 2 jours

1160 €

5 au 6 février
24 au 25 mai

20 au 21 septembre
6 au 7 décembre

Public:

Chefs de projet, développeurs, data scientists, architectes souhaitant mettre en oeuvre des solutions analytics avec hadoop.

Objectifs:

Savoir mettre en oeuvre les frameworks analytics dans un environnement hadoop.

Connaissances préalables nécessaires:

Connaissances des principes du BigData, d'un langage de programmation comme Java ou Scala ou Python.

Programme:

- Introduction : Définitions : Analytics.
Arbres de décision, de régression, régression automatique
Classifieurs, scoring.
Apprentissage supervisé, apprentissage automatique
Etapes de préparation des données.
Présentation du data munging.
- Hadoop et les outils d'analyse : Rôle des différents composants :
socle hadoop, yarn, hdfs
Frameworks analytics : Mahout, Flink, Spark ML
- Mahout : Principe de fonctionnement.
Sources de données, format de stockage des données,
Génération de recommandations, traitement, filtrage
Exemples de base : génération de recommandations,
traitement, filtrage
Présentation des algorithmes les plus courants.
Compatibilité avec Hadoop Yarn, Spark, H2O, Flink

Hadoop : analytics

- Flink** : Origine du projet, fonctionnalités
Traitement distribué de flux de données, en temps réel ou batch
APIs disponibles
Mise en oeuvre avec des programmes Java/Scala
Analyse de graphe avec l'API Gelly
- Spark MLib** : Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistance, statistiques. Support de RDD. Mise en oeuvre avec les DataFrames.
- GraphX** : Fourniture d'algorithmes, d'opérateurs simples pour des calcul statistiques sur les graphes
Travaux pratiques :
exemples d'opérations sur les graphes.