Pythagore F.D.

Apprendre à apprendre

CB033

Hadoop : développement

Durée: 3 jours

1830 €	5 au 7 mars	24 au 26 septembre
	25 au 27 juin	3 au 5 décembre

Public:

Chefs de projets, développeurs, data-scientists, et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop.

Objectifs:

Connaître les principes du framework Hadoop et savoir utiliser la technologie MapReduce pour paralléliser des calculs sur des volumes importants de données.

Connaissances préalables nécessaires:

Connaissance d'un langage de programmation objet comme Java.

Programme:

Introduction : Les fonctionnalités du framework Hadoop
Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce
Utilisation de yarn pour piloter les jobs mapreduce.

MapReduce : Principe et objectifs du modèle de programmation MapReduce.
Fonctions map() et reduce(). Couples (clés, valeurs).
Implémentation par le framework Hadoop.
Etude de la collection d'exemples.
Travaux pratiques :
Rédaction d'un premier programme et exécution avec Hadoop.

Programmation : Configuration des jobs, notion de configuration.
Les interfaces principales : mapper, reducer,
La chaîne de production : entrées, input splits, mapper, combiner, shuffle/sort, reducer, sortie.
partitioner, outputcollector, codecs, compresseurs..
Format des entrées et sorties d'un job MapReduce : InputFormat et OutputFormat.
Travaux pratiques : type personnalisés : création d'un writable spécifique. Utilisation. Contraintes.

Pythagore F.D.

Apprendre à apprendre

CB033

Hadoop : développement

Outils complémentaires	: Mise en oeuvre du cache distribué. Paramétrage d'un job : ToolRunner, transmission de propriétés. Accès à des systèmes externes : S3, hdfs, har, ... Travaux pratiques : répartition du job sur la ferme au travers de yarn.
Streaming	: Définition du streaming map/reduce. Création d'un job map/reduce en python. Répartition sur la ferme. Avantage et inconvénients. Liaisons avec des systèmes externes. Introduction au pont HadoopR Travaux pratiques : suivi d'un job en streaming.
Pig	: Présentation des pattern et best practices Map/reduce. Introduction à Pig. Caractéristiques du langage : latin. Travaux pratiques : installation/lancement de pig. Ecriture de scripts simples pig. Les fonctions de base. Ajouts de fonctions personnalisées. Les UDF. Mise en oeuvre.
Hive	: Simplification du requêtage. Etude de la syntaxe de base. Travaux pratiques : Création de tables. Ecriture de requêtes. Comparaison pig/hive.
Sécurité en environnement hadoop	: Mécanisme de gestion de l'authentification. Travaux pratiques : configuration des ACLs.