Pythagore F.D.
Apprendre à apprendre

CB040

Pig : développement de scripts

Durée: 2 jours

1320 €	8 au 9 mars	27 au 28 septembre
	7 au 8 juin	20 au 21 décembre

Public:

Chefs de projet, data scientists, développeurs souhaitant utiliser pig pour l'analyse de données

Objectifs:

Comprendre le fonctionnement de pig, savoir développer des requêtes en latin, pour effectuer des transformations sur des données,des analyses de données,intégrer des données de différents formats.

Connaissances préalables nécessaires:

Connaissance de Java ou Python, des bases Hadoop, et notions de calculs statistiques.

Programme:

Introduction : Le projet Apache Pig, fonctionnalités, versions
Présentation de Pig dans l'écosystème Hadoop.
Chaîne de fonctionnement.
Comparatif avec l'approche Hive ou Spark

Mise en oeuvre : Rappels sur les commandes HDFS
Prérequis techniques, configuration de Pig
Travaux pratiques:
Exécution : les différents modes : interactif ou batch
Principe de l'exécution de scripts Pig Latin avec Grunt

Base latin : Modèles de données avec Pig
Intégration Pig avec MapReduce
Les requêtes Latin : chargement de données, instructions
Ordres de bases :
LOAD, FOREACH, FILTER, STORE.
Travaux pratiques : création d'un ETL de base
Contrôle d'exécution

Transformations : Groupements, jointures, tris, produits cartésiens.
Transformation de base de la donnée.
Découpages. Découpages sur filtres.

Pythagore F.D.

Apprendre à apprendre

CB040

Pig : développement de scripts

Analyse de la donnée	: Echantillonages. Filtres. Rangements avec rank et dense. Calculs : min/max, sommes, moyennes, ... Ttravaux pratiques : Traitements de chaînes de caractères. Traitement de dates.
Intégration	: Formats d'entrées/sorties. Interfaçage avro, json. Travaux pratiques : chargement de données depuis HDFS vers HBase, analyse de données Pig/Hbase et restitution Json.
Extensions	: Extension du PigLatin. Création de fonctions UDF en java. Intégration dans les scripts Pig. Travaux pratiques : Utilisation de Pig Latin depuis des programmes Python Execution de programmes externes, streaming.