Pig : développement de scripts

CB040

Durée: 2 jours

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

Atelier : Exécution : les différents modes : interactif ou batch

Atelier : 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.

Atelier : 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.

Analyse de la donnée

Echantillonnages. Filtres. Rangements avec rank et dense.
Calculs : minimaux/maximaux, sommes, moyennes, ...

Atelier : traitements de chaînes de caractères. Traitement de dates.

Intégration

Formats d'entrées/sorties. Interfaçage avro, json.

Atelier : chargement de données depuis HDFS vers HBase, analyse de données Pig/Hbase et restitution Json.

Extensions

Extension de Pig/Latin.
Création de fonctions UDF en java.
Intégration dans les scripts Pig.

Atelier : utilisation de Pig Latin depuis des programmes Python

Exécution de programmes externes, streaming.