Pig, développement de scripts

Référence : **PYCB040** Durée : **2 jours** Certification : **Aucune**

CONNAISSANCES PREALABLES

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

PROFIL DES STAGIAIRES

• 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. .

CERTIFICATION PREPAREE

Aucune

METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

FORMATEUR

Consultant-Formateur expert Bigdata

METHODE D'EVALUATION DES ACQUIS

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

CONTENU DU COURS

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.

Analyse de la donnée

- Echantillonages. Filtres. Rangements avec rank et dense.
- Calculs : min/max, sommes, moyennes, ...
- 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.