Hadoop : développement

Référence : **PYCB033B** Durée : **3 jours** Certification : **Aucune**

CONNAISSANCES PREALABLES

• Connaissance d'un langage de programmation objet comme Java.

PROFIL DES STAGIAIRES

• 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.

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

• 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
• 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

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