EMGS GROUP
CONSULTING | RECRUITMENT | TRAINING

Contact
☎ 01 75 43 05 58
✉ training@emgsgroup.com

Analyse de données pour Hadoop 2.X Hortonworks avec Pig, Hive et Spark

Code formation : EMPHS
Durée : 4 jours – 28h de cours
Format : Inter-entreprise*

4 jours
28 heures de cours

**Cette formation est également disponible en « Intra-entreprise », nous contacter pour plus d'infos.*

| Description

Cette formation présente les grands outils de l'écosystème Hadoop en se focalisant plus spécifiquement sur Pig et Hive. Le principal objectif est le développement de compétences de data analyst orientées accès et traitement des données sans nécessairement avoir un fort background technique.

| Objectifs pédagogiques

- Identifier et définir les différents composants de l'écosystème Hadoop
- Appréhender l'architecture de Hadoop 2.X
- Expérimenter les outils d'exploration et d'analyse avancée de données

| Publics

Analyste, statisticien, développeur

| Pré-requis

Connaissances de base en scripting (SQL, Python, R) ou en programmation.

| Méthode pédagogique

Formation mêlant des apports théoriques à de nombreux travaux pratiques sous forme d'exercices d'application et d'analyse de uses cases métier complétés des retours d'expérience du formateur.

| Programme détaillé

Jour 1

Comprendre HADOOP 2.X

- L'architecture de Hadoop 2.X
- The Hortonworks Data Platform (HDP)

Le système de fichiers distribué HDFS

- Architecture fonctionnelle de HDFS
- Exercice d'interaction en ligne de commande avec HDFS

Alimenter HDFS en données

- Prise en main de l'outil Flume
- Prise en main de l'outil Sqoop
- Application de ces deux outils d'import et d'export des données

Le framework MAPREDUCE

- Architecture et fonctionnement général de MapReduce
- Exemples d'utilisation d'un job MapReduce
- Présentation de Hadoop Streaming

Jour 2

Introduction à PIG

- Types et mots-clés dans Pig
- Exploration de données avec Pig

Programmation PIG avancée

- Mots-clés et fonctionnalités avancées dans Pig
- Jointures dans Pig
- Astuces d'optimisation de scripts Pig
- Analyse de cas d'usages métier divers avec Pig

Jour 3

Programmation HIVE

- Types et mots-clés dans Hive
- Concept de table et base de données dans Hive
- Présentation et explication des types de jointures
- Démonstration de jointures
- Analyse de cas d'usages métier

Utiliser HCATALOG

- Fonctionnement et utilisation de HCatalog
- Démonstration du fonctionnement de HCatalog

Jour 4

Programmation HIVE avancée

- Les vues dans Hive
- Les différents formats de stockage des tables Hive
- Optimisation de scripts Hive
- Illustration des fonctions avancées

HADOOP 2.X et YARN

- Architecture de YARN
- Démonstration d'une application YARN

APACHE SPARK

- Introduction à Spark
- Programmation Spark (RDD, programmation fonctionnelle)
- Ecriture d'un job Spark en Python
- Spark SQL et les DataFrames
- Utilisation de Spark SQL et des DataFrames sur des tables Hive et des fichiers HDFS

Créer et utiliser un workflow Oozie

- Workflow et coordinateur Oozie
- Actions possibles avec Oozie