

Formation IBM InfoSphere

BigInsights : les bases

Objectifs de la formation IBM InfoSphere BigInsights

Cette formation "Les bases d'IBM Infosphere BigInsights" présente aux participants une vue d'ensemble de la stratégie Big Data d'IBM et de la solution BigInsights en tant que plateforme pour gérer et obtenir des informations de données. Lors de ce cours les participants vont découvrir la valeur ajoutée à BigInsights incluant Big SQL, BigSheets et Big R. Ils verront ensuite en quoi IBM Open Platform (IOP) avec Apache Hadoop est une plateforme collaborative permettant à des solutions Big Data d'être développées sur un ensemble commun de technologies Apache Hadoop.

Il est également réalisé une présentation détaillée des composants principaux du noyau ODP, à savoir Apache Hadoop (y compris HDFS, YARN et MapReduce) et Apache Ambari, ainsi que leurs traitements.

À l'issue de cette formation, les participants seront capables de :

- Comprendre les objectifs des big data et savoir en quoi c'est important
- Lister les sources de données
- Décrire la solution IBM BigInsights
- Utiliser les nombreux outils IBM BigInsights incluant Big SQL, BigSheets, Big R, Jaql et AQL
- Lister et décrire les composants majeurs d'une pile open-source Apache Hadoop et des approches de l'Open Data Foundation.
- Gérer et contrôler les clusters Hadoop avec Apache Ambari et les composants associés
- Explorer Hadoop Distributed File System (HDFS) en lançant les commandes Hadoop

Capital Social: 50000 DT **MF:** 1425253/M/A/M/000 **RC:** B91211472015

Tél. / Fax.: +216 73362 100 **Email:** contact@itgate-training.com **Web:** www.itgate-training.com

Adresse : 12 Rue Abdelkadeur Daghbir - Hammam Sousse 4011 – Tunisie

- Comprendre les différences entre Hadoop 1 (avec MapReduce 1) et Hadoop 2 (avec YARN et MapReduce 2)
- Créer et exécuter des tâches basiques MapReduce en ligne de commande
- Expliquer comment Spark s'intègre dans l'écosystème Hadoop
- Exécuter des algorithmes itératifs avec Spark RDD.
- Expliquer le rôle de coordination, de gestion, et de gouvernance dans l'écosystème Hadoop en utilisant Apache Zookeeper, Apache Slider, et Apache Knox.
- Explorer les méthodes communes pour assurer le mouvement des données
- Configurer Flume pour le chargement de données et de fichiers logs
- Déplacer les données dans HDFS depuis des bases relationnelles avec Sqoop
- Comprendre quels formats de stockage de données utiliser (flat files, CSV/delimited, Avro/Sequence files, Parquet, etc.)
- Etudier les différences entre les langages de programmation open-source généralement utilisés avec Hadoop (Pig, Hive) et pour la Data Science (Python, R)
- Requêter des données depuis Hive
- Assurer un accès aléatoire sur des données stockées dans HBase
- Explorer les concepts avancés, incluant Oozie et Solr.

À qui s'adresse cette formation ?

Public :

Cette formation est destinée aux professionnels intéressés par le Big Data et la solution IBM BigInsight. Elle concerne plus précisément les ingénieurs data, les data scientists, les développeurs et les administrateurs souhaitant en savoir plus sur la plateforme IBM Open Platform avec Apache Hadoop.

Prérequis :

Afin de profiter pleinement de ce cours, il est recommandé d'avoir des connaissances sur Linux.

Contenu du cours IBM InfoSphere BigInsights

Introduction au Big Data

Studio :

Configuration de l'environnement lab

Présentation d'IBM BigInsights

studio :

Prise en main d'IBM BigInsights

IBM BigInsights for Analysts

studio:

Utilisation de Big SQL et BigSheets

IBM BigInsights for Data Scientist

studio:

Analyse des données avec Big R, Jaql et AQL

IBM BigInsights for Enterprise Management

IBM Open Platform avec Apache Hadoop

studio:

Explorer le HDFS

Capital Social: 50000 DT **MF:** 1425253/M/A/M/000 **RC:** B91211472015

Tél. / Fax.: +216 73362 100 **Email:** contact@itgate-training.com **Web:** www.itgate-training.com

Adresse : 12 Rue Abdelkadeur Daghrrir - Hammam Sousse 4011 – Tunisie

Apache Ambari

studio:

Gestion des clusters Hadoop avec Apache Ambari

Systeme de fichiers distribués Hadoop

studio:

Accès aux fichiers & commandes de base avec HDFS

MapReduce et Yarn

Introduction à MapReduce basé sur MR1

Limites de MR1

YARN et MR2

studio:

Création et codage d'une tâche MapReduce simple (peut-être un deuxième exercice plus complexe)

Apache Spark

studio:

Utilisation du RDD de Spark dans une tâche Spark

Coordination, gestion et gouvernance

studio:

Apache ZooKeeper, Apache Slider, Apache Knox

Déplacement des données

studio:

Déplacement de données dans Hadoop avec Flume et Sqoop

Stockage et accès aux données

Représentation de données : CSV, XML, JSON et YAML

Langages de programmation Open Source : Pig, Hive et autres (R, Python, etc.)

NoSQL Concepts

Accès aux données Hadoop à l'aide de Hive

studio:

Exécution d'opérations CRUD à l'aide du shell HBase

Interrogation de données Hadoop à l'aide de Hive

studio:

Utilisation de Hive pour accéder aux données Hadoop / HBase

Rubriques avancées

Contrôle des workflows de travail avec Oozie

Recherche à l'aide d'Apache Solr Aucun exercice de laboratoire