Big Data - Mise en oeuvre pratique d'une solution complète d'analyse des données

Collecte et analyse de données massives

Présentiel ou en classe à distance

4 jours (28 h)
+ activité à distance

Prix inter : 2.890,00 € HT
Forfait intra : 9.850,00 € HT

Réf.: BD550



Comprendre les besoins métier pour livrer aux décideurs des indicateurs fiables et pertinents, tel est le rôle attendu des spécialistes Big Data. Précisément conçu pour leur apprendre à mettre en oeuvre une solution de Big Data en environnement Hadoop, solution phare pour les traitements Big Data, ce programme reprend le cheminement logique d'un projet d'analyse de données. De leur collecte initiale à la mise en place de solutions de stockage spécifiques permettant d'organiser un très grand volume d'information, à la réalisation de scripts Pig et Hive qui, convertis en tâches MapReduce, permettent d'agréger et de filtrer les données pour finalement les analyser, tous les aspects seront abordés.

A qui s'adresse cette formation ?

- Chefs de projet
- Data Scientists, Data Analysts
- Développeurs
- Analystes et statisticien
- Toute personne en charge de la mise en oeuvre opérationnelle d'un projet Big Data en environnement Hadoop

- Avoir suivi la formation "Big Data - Les fondamentaux de l'analyse de données" (BD540)
- La connaissance d'un langage de programmation et du langage SQL est un plus pour suivre cette formation
- Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !

Programme

Avant

- un quiz de consolidation des prérequis

Pendant

1 - Introduction

- Objectifs
- Schématisation du projet

- Écosystème et stack technologique
- Résultats attendus

2 - Ingestion de données massives

- Description
- Caractéristiques clés des outils d'ingestion
- Solutions (packagées, cloud computing et Hadoop)
- Focus Apache NIFI
- Focus Apache KAFKA
- Ingestion de données en streaming NIFI on KAFKA
- Réalisation d'un workflow NIFI d'ingestion de donnée streaming dans HDFS

3 - Traitement de données Big Data en batch

- Diagramme de fonctionnement
- Solutions logicielles associées (MapReduce, Hive, Spark, ...)
- Big Data Batch scripting
- Data Warehousing Big Data
- Création de tables partitionnées, clusterisées et/ou transactionnelles
- Écriture des scripts d'updates des différentes tables
- Analytics Big Data

4 - Traitement avancé Big Data : l'apprentissage machine

- L'écosystème Spark
- Schéma général de création d'un modèle de ML
- Création d'un modèle de clusterisation de données
- Création d'un modèle d'analyse prédictive supervisé
- Enrichissement d'un DWH avec Spark
- Sauvegarde d'un modèle d'apprentissage machine
- Application d'un modèle ML

5 - Stockage de données distribuées

- Principes des bases de donnes distribuées
- Solutions (NoSQL, NewSQL)
- Création d'une base de données distribuée
- Ingestion de données dans une base de données distribuées
- Interrogation d'une base de données distribuées

6 - Automatisation de chaîne de traitement Batch

- L'orchestrateur Oozie
- Ordonnancement de scripts HIVE
- Combinaison avec des scripts SPARK
- Création d'un worklow Oozie complet

7 - Traitement de données massives en flux (streaming)

- Principe de fonctionnement
- Solutions logicielles
- Le prétraitement de données en streaming
- L'inscription de streams à un Hub Streaming
- La consommation de streams auprès d'un Hub
- Le traitement avancé de données en flux (machine learning)

8 - Mise en oeuvre dans une architecture Big Data

- Les approches standards
- Réalisation d'une solution complète de traitement de données type Lamda
- Réalisation d'une solution complète de traitement de données type Kappa

Après

- Un quiz pédagogique pour évaluer vos acquis et approfondir les sujets de votre choix
- Des vidéocasts pour revenir sur les points clés de la formation
- Des vidéos-tutos pour vous accompagner dans l'utilisation des outils du Big Data

Les objectifs de la formation

- Disposer des compétences techniques nécessaires à la mise en oeuvre d'analyses Big Data
- Comprendre le cadre juridique du stockage et de l'analyse de données
- Savoir utiliser des outils de collecte Open Source
- Être en mesure de choisir la bonne solution de stockage de données au regard des spécificités d'un projet (OLAP, NoSQL, graph)
- Explorer la boite à outils technologique que constitue Hadoop et son écosystème et savoir comment utiliser chaque brique (MapReduce, HIVE, SPARK...)
- Apprendre à analyser les résultats et comprendre la signification des données extraites

Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Les points forts de la formation

- Une formation très opérationnelle durant laquelle s'alternent les phases d'apports théoriques, d'échanges, de partage d'expériences.
- Les participants réalisent sur un cas concret une chaîne de traitement Big Data (Batch et Streaming).
- Ils travaillent dans un environnement Cloud sur une des distributions Hadoop les plus utilisées.
- Les contenus digitaux mis à disposition des stagiaires avant et après la formation renforcent l'efficacité pédagogique du programme et garantissent un bénéfice durable de l'action de formation.
- 75% des participants à cette formation se sont déclarés satisfaits ou très satisfaits au cours des 12 derniers mois.

Evaluation

Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.

Dates et villes 2027 - Référence BD550

Dernières places disponibles Session garantie

A distance

du 15 mars au 18 mars	du 20 sept. au 23 sept.
du 21 juin au 24 juin	du 15 nov. au 18 nov.

Paris

du 15 mars au 18 mars	du 20 sept. au 23 sept.
du 21 juin au 24 juin	du 15 nov. au 18 nov.