Date de publication : 26/09/2025

Formation : Hadoop Cloudera développeur, préparation à la certification (CCA175)

Formation pratique - 4j - 28h00 - Réf. HDZ

Cette formation vous apprendra les concepts clés et l'expertise nécessaire pour intégrer et enregistrer les données dans un cluster Hadoop avec les techniques et les outils plus récents. Elle prépare à la certification "CCA Spark and Hadoop developer".

Objectifs pédagogiques

À l'issue de la formation, le participant sera en mesure de :

- Découvrir l'écosystème Hadoop
- Comprendre le système de fichiers distribué HDFS et maîtriser le traitement MapReduce et l'écriture de code
- Connaître les bonnes pratiques de développement et d'implémentation des algorithmes courants
- Optimiser les configurations et améliorer les performances
- Utiliser Hive, Pig, Flume, Mahout et Sqoop pour les projets de l'écosystème Hadoop
- Préparer la certification Cloudera

Public concerné

Chefs de projets, développeurs, data scientists, et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop.

Prérequis

Connaissances de base dans un langage de programmation objet.

Vérifiez que vous avez les prérequis nécessaires pour profiter pleinement de cette formation en faisant ce test.

Certification

À la suite de la formation, il sera possible de passer l'examen « Cloudera Certified Associate Spark and Hadoop Developer (CCA175) ». Cet examen se déroule en dehors du temps de la formation. L'objectif est de devenir expert certifié Cloudera dans son entreprise. Inscriptions sur www.examslocal.com.

PARTICIPANTS

Chefs de projets, développeurs, data scientists, et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop.

PRÉREQUIS

Connaissances de base dans un langage de programmation objet.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques...
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

Méthodes et moyens pédagogiques

Méthodes pédagogiques

Cette formation big data comprend 50% de travaux pratiques sur les 4 jours de formation.

Modalités d'évaluation

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…

Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

Programme de la formation

1 Hadoop, HDFS et traitement distribué sur un cluster Hadoop

- Introduction générale à Hadoop et à son écosystème.
- Traitement de données.
- HDFS : le système de fichiers Hadoop.
- Les composants d'un cluster Hadoop.
- L'architecture d'HDFS. Utiliser HDFS.
- L'architecture de YARN et travailler avec YARN.

2 Les bases de Spark

- Introduction à Spark.
- Démarrer et utiliser la console Spark.
- Introduction aux Datasets et DataFrames Spark.
- Les opérations sur les DataFrames.

3 Manipulation des DataFrames, des schémas, analyse des données avec requête

- Créer des DataFrames depuis diverses sources de données.
- Sauvegarder des DataFrames. Les schémas des DataFrames.
- Exécution gloutonne et paresseuse de Spark.
- Requêter des DataFrames avec des expressions sur les colonnes nommées.
- Les requêtes de groupement et d'agrégation.
- Les jointures.

4 Les RDD et requêtage de tables et de vues avec Spark SQL

- Structure fondamentale de Spark.
- Transformer les données avec des Resilient Distributed Dataset (RDD).
- Agrégation des données avec les RDD de paires.
- Requêter des tables en Spark en utilisant SQL.
- Requêter des fichiers et des vues.
- L'API catalogue de Spark.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les formations pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque formation ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le participant a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Pour toute question ou besoin relatif à l'accessibilité, vous pouvez joindre notre équipe PSH par e-mail à l'adresse psh-accueil@orsys.fr.

5 Travailler avec Spark

- Travailler avec les datasets Spark en Scala. Les différences entre datasets et DataFrames.
- Créer, charger et sauvegarder des datasets. Les opérations sur les datasets.
- Écrire, configurer et lancer des applications Spark.
- Écrire une application Spark. Compiler et lancer une application. Le mode de déploiement d'une application.
- L'interface utilisateur web des applications Spark. Configurer les propriétés d'une application.
- Le traitement distribué avec Spark. Rappels sur les fonctionnements de Spark avec YARN.
- Le partitionnement des données dans les RDD, dans les requêtes, jobs, étapes et tâches.

6 Persistance de la donnée distribuée

- La persistance des DataFrames et des datasets.
- Les niveaux de persistances.
- Les RDD persistés

7 Les algorithmes itératifs avec Spark et introduction à Spark streaming

- D'autres cas d'usages courants de Spark.
- Les algorithmes itératifs en Spark. Machine learning avec Spark.
- Introduction à Spark streaming. Créer des streaming DataFrames.
- Transformer des DataFrames. Exécuter des requêtes de streaming.

8 Structured streaming avec Kafka et opérations sur des streaming DataFrames

- Introduction. Recevoir et envoyer des messages Kafka.
- Agrégation et jointure sur des streaming DataFrames.