Formation certifiante Hadoop avec Spark pour Développeurs de Cloudera (CCA Spark and Hadoop Developer)

https://training.xebia.fr/formation-data/formation-certifiante-hadoop-avec-spark-pour-developpeurs/

Certification Hadoop avec Spark pour Développeurs de Cloudera (CCA Spark and Hadoop developer)

Avec ***Bruno Bouchahoua***

4 jours, soit 28 heures

Ce cours pratique de 4 jours fournira aux stagiaires 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.

Les stagiaires utiliseront des projets tels que Spark, Hive, Flume, Sqoop et Impala afin de bénéficier de la meilleure préparation possible pour faire face aux défis quotidiens auxquels sont confrontés les **développeurs Hadoop**. Les participants apprendront à identifier et à utiliser les outils appropriés à chaque situation.

Apprendre comment importer des données dans votre « cluster » Apache Hadoop et le transformer

avec Spark, Hive, Flume, Sqoop, Impala, et d’autres outils de l’écosystème Hadoop.

Programme

Jour 1

Introduction

- A propos du cours
- Cloudera
- Logistique du cours
- Présentations

Apache Hadoop et son écosystème

- Introduction à Hadoop
- Stockage et ingestion des données
- Processing des données
- Analyse des données et exploration
- Autres outils de l’écosystème

Stockage de fichiers sur Hadoop

- Composants principaux d’un cluster
- Architecture d’ HDFS
- Utilisation d’ HDFS
- Format de fichier sur Hadoop

Processing des données sur un cluster Hadoop

- Architecture de YARN
- Travailler avec YARN

Importer les données d’une base de données relationnelle sur Hadoop

- Présentation de SQOOP
- Importer des données avec SQOOP
- Options d’import
- Exporter des données

Apache Spark les bases

- Qu’est ce que Apache Spark ?
- Utiliser le Shell de Spark
- RDD
- Programmation fonctionnelle au sein de Spark

Jour 2

Travailler avec les RDD

- Créer des RDD
- Opérations principales avec les RDD

Agrégation des données avec les pair RDD

- Key-value Pair RDD
- Map Reduce
- Autres opérations avec les Pair-RDD

Ecrire et exécuter des applications Spark

- Spark Shell versus Spark Application
- Création du Spark Context
- Construction d’un application Spark
- Lancement d’une application Spark
- Web UI relatives à Spark

Configuration des applications Spark

- Propriétés de configuration de Spark
- Gestion des log

Exécution distribuée

- Spark en exécution sur un cluster
- Partition des RDD
- Partition des RDD basés sur des fichiers
- “Data Locality” sur HDFS
- Exécution des opérations en parallèle

Jour 3

Persistance des RDD

- Cycle de vie d’un RDD
- Persistance d’un RDD
- Persistance distribuée

Traitements communs pour le processing de données avec Sparks

- Cas d’utilisation de Spark
- Algorithmes itératifs
- Machine learning
- K-means

DataFrame et Apache Spark SQL

- Apache Spark SQL et le SQL Context
- Création des Dataframes
- Transformer et requêter un Dataframe
- Persister un Dataframe
- Dataframes et RDD
- Comparaison entre Spark SQL, Impala et Hive On Spark

Traitement des messages avec Apache Kafka

- Qu’est ce que Apache Kafka ?
- “Scaler” Apache Kafka
- Architecture d’un cluster Kafka
- Outils en ligne de commande d’Apache Kafka

Jour 4

Récupération des événements avec Apache Flume

- Qu’est ce que Apache Flume ?

- Architecture
- Sources
- Sinks
- Canaux
- Configuration

Intégration entre Apache Flume et Apache Kafka

- Présentation
- Cas d’utilisation
- Configuration

Apache Spark Streaming: Introduction DStream

- Présentation de Apache Spark Streaming
- Exemple de cas d’utilisation temps réel
- DStreams
- Développement d’application temps réel

Apache Spark Streaming: Processing multi-batch

- Opérations Multi batch
- Découpage par période
- Maintien d’un état
- Déplacement au travers d’une fenêtre de temps

Apache Spark Streaming: Sources de données

- Sources de données pour le streaming
- Flume et Kafka comme source de données
- Source de données Kafka

Objectifs

Les stagiaires, à l’issue de la formation, sauront :

- Identifier et utiliser les outils appropriés à chaque situation dans un écosystème Hadoop
- Utiliser Apache Spark et l’intégrer dans l’écosystème Hadoop
- Utiliser Hive, Impala, Flume et Sqoop

Méthodes pédagogiques

Les participants apprendront **Apache Spark** et comment l’intégrer dans l’écosystème Hadoop au travers d’échanges avec le formateur ainsi qu’en réalisant des exercices pratiques sur les sujets suivants :

- Comment les données sont distribuées, stockées et enregistrées dans un cluster Hadoop ?
- Comment utiliser Sqoop et Flume pour intégrer les données ?
- Comment enregistrer des données distribuées avec Apache Spark ?
- Comment modéliser des données structurées en tant que tableau dans Impala et Hive ?
- Comment choisir le meilleur format de stockage de données pour différents patterns d'utilisation de données ?
- Les meilleures pratiques pour le stockage de données.

Voici un exemple d'exercice pratique:

Vous êtes embauché par une société fictive LOUDACRE spécialisée dans fourniture de réseau mobile. Votre rôle va être d'accompagner cette société dans sa transformation vers l'adoption du big data. Tout son système d'information existant doit être migré vers un cluster Hadoop pour lui permettre de supporter sa forte croissance et son volume important de données à traiter.

Technologiques de l'écosystème Hadoop : Impala, HDFS, Hue, Yarn, Sqoop, Spark, Spark Streaming, Spark Dataframe, Apache Kafka, Apache Flume.

Tarif HT

Inter : 2 995 €

Prochaines dates de formation

21 – 24 août
22 – 25 octobre
27 – 30 novembre

Je m'inscris

info@xebia-training.fr

Programme PDF

Imprimer le programme

Public visé

Cette formation est prévue pour des développeurs et des ingénieurs qui ont une expérience de programmation.

Prérequis

Les exemples Apache Spark et les exercices de « hands-on » sont présentés avec Scala et Python, donc il faut être à l'aise pour programmer dans l'un de ces langages. Avoir une connaissance de base avec les lignes de commande Linux est requis. Avoir une connaissance de base de SQL est utile. **Aucune expérience préalable avec Hadoop n'est nécessaire.**
Les postes de travail et les logiciels nécessaires au bon déroulement de la formation sont fournis par Xebia

Certification

A la suite de la formation, les stagiaires auront la possibilité de passer l'examen Certification « CCA Spark and Hadoop Developer » de Cloudera. Cet examen se déroule en dehors du temps de la formation. Ils deviendront alors des experts certifiés Cloudera dans leur entreprise.

Validation

À la fin de cette formation, les stagiaires recevront une attestation de présence.