www.bigdataschool.ru

ШКОЛА БОЛЬШИХ ДАННЫХ

SPARK: Анализ данных с Apache Spark

Длительность: 24 ак. часов

О курсе

3х-дневный интенсивный практический курс для разработчиков **Apache Spark**, дата инженеров и аналитиков данных, **Data Scientists** и других специалистов **Big Data**, которые используют **Spark SQL**, потоковую обработку **Spark Streaming**, машинное обучение **MLLib** и построение графов **Spark GraphX**.

Аудитория

Разработчики Big Data, дата инженеры и аналитики данных, Data Scientists и другие специалисты по большим данным, которые хотят получить опыт настройки и использования компонентов **Apache Spark**: **Spark Core**, **Spark SQL**, **Spark Streaming**, **Spark MLLib** и **Spark GraphX**.

Соотношение теории к практике 40/60

Предварительная подготовка

- **Unix** - уверенное владение командной строкой bash, знание основных команд, принципов работы файловой системы
- **SQL** - написание запросов среднего уровня сложности
- **Python** - опыт программирования от 2 лет
- **Экосистема Hadoop** - знание основных компонент, понимание их ролей и взаимосвязей

sales@bigdataschool.ru
+7 (915) 307-00-74

Программа курса

1. **Обзор Apache Spark, знакомство со Spark RDD и Dataframes**
   Архитектура Spark. Принципы работы **Resilient Distributed Dataset** (**Spark RDD**)
   Обзор компонентов **Spark** и их назначения
   Low Level API, использование **Resilient Distributed Dataset**
   Structured API и основная абстракция **Spark** - **Dataframe**

2. **Apache Spark SQL**
   Получение данных из **SQL**-источников и обработка данных с помощью **Spark SQL**
   Отправка данных в **SQL СУБД** и работа с **Hive QL**
   Spark SQL и **Hadoop**

3. **Производительность и параллелизм в Apache Spark**
   Планы выполнения запроса: логические и физические
   Конфигурирование **Apache Spark**

4. **Spark Streaming**
   Разница работы в режимах **OLAP** и **OLTP**. Основной **workflow**
   Виды **Spark Streams**. Особенности исполнения **streaming** кода
   Checkpoint в **Spark Streaming**

5. **GraphX и MLLib**
   Задачи графов в программировании. Место графов в модели распределенных вычислений
   Представление графов в **GraphX**. Операции с графами
   Задачи машинного обучения и проблематика больших данных
   Основные возможности **Spark MLLib**

6. **Обработка слабоструктурированных данных**
   Работа с **JSON** файлами и строками
   Обработка информации, представленной в виде **XML**

sales@bigdataschool.ru
+7 (915) 307-00-74