Курс: Анализ больших данных с Apache Spark

Длительность: 24 ак. часов

О курсе

3х дневный интенсивный практический тренинг по знакомству с платформой для распределенной обработки больших данных **Apache Spark**. В ходе лекций и лабораторных работы вы научитесь настраивать кластер **Apache Spark** для запуска задач на **Scala** и **R** при обработке больших массивов неструктурированных данных, применяя алгоритмы **машинного обучения** встроенных библиотек **Spark MLLib**; поймете разницу использования различных форматов хранения данных и использования **RDD** , **dataframes** и **datasets**; обращаться к данным с использование **Spark SQL** или **Hive QL**; настраивать и анализировать данные в потоке **Spark Streaming**; интегрировать компоненты **Apache Spark** с другими компонентами экосистем **Hadoop**.

Аудитория

Специалисты, администраторы, аналитики данных желающие получить опыт настройки и использования компонентов **Apache Spark** (**Spark SQL**, **MLLib**, **Spark Streaming**, **Spark GraphX**)

Соотношение теории к практике 40/60

Предварительная подготовка

- Начальный опыт работы в **Unix/SQL,** текстовый редактор **vi**
- Начальный опыт программирования (**Scala**/Python/Java) **(приветствуется)**

Программа курса

1. **Введение в Apache Spark**
   Сравнение **Hadoop и Spark**
   Сравнение **Batch**, **Real-Time** и **in-Memory** процессинг
   Особенности **Apache Spark**
   Компоненты **Apache Spark** экосистемы

2. **Введение в RDD - Resilient Distributed Dataset**
   Что такое **RDD**
   Особенности использования **RDD, RDD lineage**
   Трансформация в **Spark RDD**
   Lazy evaluation и отказоустойчивость в **Spark**
   Использование Persistence RDD в памяти и на диске
   Использование key-value пар (**ReduceByKey**,**CountByKey**,**SortByKey**,**AggregateByKey**)
   Интеграция **Hadoop** с **Spark**

3. **Запуск задач в Apache Spark**

sales@bigdataschool.ru
+7 (985) 162-29-63

Знакомство с **Spark-shell**
Выполнение задач в **Apache Spark**
Написание программ в **Apache Spark**
Чтение данных с локальной файловой системы и **HDFS**
Зависимости(**Dependencies**)
Кэширование данных в **Apache Spark**
Отказоустойчивость (**Fault Tolerance**)

4. **SparkSQL, DataFrames, DataSet**
   Альтернатива RDDs
   Сравнение **DataFrame**, **DataSet** и **SQL API**
   Введение в **SparkSQL,** пользовательские функции в **Spark SQL**
   Использование **DataFrames** и **DataSet, DataSets** вместо **RDD**
   Простые запросы, фильтрация и аггрегация **DataFrames**
   Объединение (**JOIN**) **DataFrames**
   Интеграция **Hive** и **Spark: Hive** запросы в **Spark,** создание **Hive** контекста, запись **Dataframe** в **Hive**

5. **Управление ресурсами в кластере Apache Spark**
   Архитектура **Apache Spark**
   Особенности управления ресурсами в автономном режиме кластера (**Standalone**)
   Особенности управления ресурсами в режиме **Hadoop** кластера с **YARN**
   Динамическое распределение ресурсов **Dynamic Resource Allocation**
   Оптимизация **Apache Spark:** использование разделов (**partition hash**,**range**,**map**, **static**), управление расписанием (**dynamic**, **fair scheduler**), использование переменных (**shared**, **broadcast**) и аккумуляторов (**accumulators**).
   Использование **Catalyst Optimizer** для оптимизации исполнения запросов
   Project Tungsten - Оптимизация управления памятью и кэшом **CPU**

6. **Машинное обучение(Machine Learning) в Apache Spark**
   Введение в **Machine Learning** с использованием **MLLib**
   Алгоритм линейной регрессии (**Linear Regression**)
   Деревья решений (**Decision Trees**)
   Случайные леса (**Random Forest**)
   Использование **DataFrames** с **MLLib**

7. **Потоковая обработка (Streaming) в Apache Spark**
   Потоковая обработка данных для аналитики больших данных
   Особенности реализации потоковой обработки данных в **Apache Spark**
   Основные концепции потоковой обработки
   Аггрегированные и не аггрегированные запросы
   Обработка событий **Event Time**, **Window** и **Late Events (**скользящее окно событий**)**
   Поддержка последних событий (**Late Events**) в потоковой обработке данных в **Apache Spark**
   Режимы работы **Apache Spark** с потоковыми данными

8. **Введение в GraphX**
   GraphX и **Pregel**
   Поиск в ширину **(Breadth-First-Search)** с использование **GraphX**