

Фреймворк Apache Spark для разработчиков: продвинутый уровень

Код: EAS-024

Длительность: 24 ч.

Описание:

Тренинг даёт подробное представление о внутреннем устройстве и функционировании фреймворка Apache Spark – как Spark Core (RDD), так и Spark SQL, Spark Streaming и Spark Structured Streaming. Рассматриваются механизмы запуска компонентов кластера Spark под управлением разных менеджеров кластеров, управление выделением ресурсов (в первую очередь – памяти), механизмы работы планировщиков. Подробно исследуются преимущества формата внутреннего представления Tungsten и работы оптимизатора Catalyst.

Цели:

- Понимать внутреннее устройство Spark;
- Понимать механизмы запуска и настройки компонентов Spark на разных кластерах (Standalone, YARN, Mesos);
- Научиться оптимизировать программы, обрабатывающие RDD;
- Понимать механизмы оптимизации программ на Spark SQL;
- Понимать механизмы оптимизации программ, обрабатывающих потоковые данные, как в микропакетах, так и в Dataset.

Разбираемые темы:

- Внутренняя архитектура Spark, Spark Runtime Environment
- Настройка Spark Context, SparkConf
- Внутреннее устройство RDD, логический план
- Лучшие практики программирования с RDD
- Физический план: работы, стадии, задачи

- Планировщики и выполнение физического плана
- Настройка памяти, сериализация, кеширование, сборка мусора
- Datasource API, внутреннее представление данных Tungsten, форматы файлов
- Оптимизатор Catalyst
- Микропакетный Spark Streaming: получение и выдача данных
- Structured Streaming: получение и выдача данных

Целевая аудитория:

Разработчики, архитекторы

Предварительная подготовка - общее:

Опыт разработки на Java или Scala под Apache Spark от 3 месяцев.

Рекомендуемые дополнительные материалы, источники:

1. Mike Frampton. Mastering Apache Spark. – Packt, 2015. – 476 pp.
2. Sandy Ryza et al. Advanced Analytics with Spark– O'Reilly, 2015. –261 pp.
3. Sumit Gupta. Learning Real-time Processing with Spark Streaming. – Packt, 2015. – 271 pp.
4. X. Карау и др. Изучаем Spark: молниеносный анализ данных. – М.: ДМК Пресс, 2015. – 304 с.