

## Программа курса «CORS: Core Spark - основы для разработчиков»

**О курсе:** 4-дневный практический курс для разработчиков Apache Spark, дата инженеров, аналитиков данных, Data Scientist'ов и других специалистов Big Data, которые используют или планируют использовать Spark для обработки и анализа больших данных.

### Аудитория:

- разработчики Big Data;
- дата инженеры;
- аналитики данных;
- Data Scientist;
- другие специалисты по большим данным, которые хотят получить опыт настройки и использования базовых компонентов Apache Spark: Low Level API, Structured API и Spark SQL.

### Уровень подготовки:

- Опыт работы в Unix/SQL;
- Уверенный опыт программирования (Python/Scala);
- Начальный опыт в экосистеме Hadoop.

**Продолжительность курса:** 16 академических часов, 4 дня по 4 ак. часов дистанционно

## Содержание программы

### 1. Обзор Apache Spark

- Архитектура Spark
- Обзор компонентов Spark и их назначения
- Spark API и разработка программ.

### 2. Основные абстракции — знакомство со Spark RDD и Dataframe

- Low Level API, использование Resilient Distributed Dataset
- Structured API и основная абстракция Spark – Dataframe
- Основные трансформации и действия.

### 3. Работа с источниками данных

- Ввод и вывод данных в Spark
- Работа с файлами и базами данных
- Совместная работа с Apache Hive

### 4. Apache Spark SQL

- Получение данных из SQL-источников
- Обработка данных с помощью Spark SQL
- Интероперабельность Dataframe и SQL

### 5. Основы настройки и конфигурирования

- Знакомство с основными Shell
- Запуск приложений с помощью spark-submit
- Конфигурирование Spark
- Планы запросов и Web GUI