Программа курса
«SPOT: Потоковая обработка в Apache Spark»

О курсе: 2х-дневный практический курс для разработчиков Apache Spark, дата инженеров и аналитиков данных, Data Scientist'ов и других специалистов Big Data, которые используют или планируют использовать Spark для обработки и анализа больших данных

Аудитория: Практический курс по потоковой обработке с использованием Спарк рассчитан на разработчиков Big Data, дата инженеров и аналитиков данных, Data Scientist'ов и других специалистов по большим данным, которые хотят получить опыт настройки и использования механизмов потоковой обработки с разными видами источников данных и нюансами практического использования возможностей Structured Streaming.

Уровень подготовки:

- Опыт работы в Unix/SQL;
- Начальный опыт программирования (Python/Java);
- Знания в объеме, аналогичном курсу Core Spark
- Начальный опыт в экосистеме Hadoop
- Базовые знания Kafka

Продолжительность курса: 16 академических часов, 2 дня по 8 ак. часов дистанционно

Содержание программы

1. Введение в потоковую обработку

- Потоковая и пакетная обработка данных
- Особенности потоковой обработки
- Надежность и потоковая обработка.

2. Потоковая обработка в Apache Spark

- Два вида потоков (на основе RDD и Dataframe)
- Парадигма потоковой обработки в Structured Streaming
- Источники (sources и sink).

3. Совместное использование Batch и Streaming

- Трансформации и действия в Apache Spark
- Объединение данных в Spark (join)
- Особенности использования трансформаций при работе с потоковыми данными

4. Источники потоковых данных

- Файловый источник данных
- Apache Kafka как источник данных
- Другие источники потоковых данных

5. Обеспечение надежности потоковой обработки в Apache Spark

- Механизм checkpoint в Apache Spark
- Настройка streaming checkpoint