

Программа курса

«FLINK: Поточковая обработка данных с помощью Apache Flink»

О курсе: 2-хдневный курс обучения по Flink позволит вам получить и систематизировать знания по использованию Flink для построения систем потоковой обработки данных. Курс нацелен на вопросы использования возможностей Apache Flink для обработки прежде всего потоковых данных, включая нюансы извлечения меток времени и генерации watermarks, лучшие практики (best practices) использования механизмов временных окон, работы с локальным состоянием, вопросы эффективности и параллелизма обработки, тонкости работы с источниками данных и многое другое.

Аудитория: Дата-инженеры и специалисты по данным, которые хотят изучить необходимый теоретический минимум по Apache Flink, чтобы понять основные концепции фреймворка, stateful преобразований и доставки данных из сторонних систем, а также узнать особенности подхода Flink к построению масштабируемых надежных конвейеров обработки потоковых данных.

Уровень подготовки: Уверенное знание базовых команд Linux (опыт работы с командной строкой), знания языка Scala (способность создания не сложных ООР приложений).

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

Содержание программы

1. Введение и архитектура Apache Flink

- Поточковая и Batch обработка
- Основные понятия и особенности потоковой обработки
- Встречаем Flink: компоненты и первое приложение
- Основные API

2. Datastream API: Базовые трансформации и работа с источниками данных

- Основные операторы (трансформации)
- Роллинг агрегации и reduce
- Работа с источниками (Apache Kafka, файловые источники)

3. Datastream API: Работа с учетом времени

- Архитектура Apache Flink
- Event Time Processing
- Оконные операции и джойны

4. Datastream API: Работа с состоянием

- Низкоуровневые «process» функции и управление состоянием
- Checkpoints, savepoints

5. Другие API, языки и системы

- Flink Table API и SQL
- Flink и python
- Сравнение Flink с Kafka Streams и Spark Structured Streaming