



О курсе

Обработка данных в режиме
реального времени

Преподаватель

Артем Гогин

Sberbank, Data engineer

Цель курса

Понимать, какие есть виды обработки потоковых данных

Уметь использовать Spark Streaming

Уметь работать с Kafka через консоль и Spark

Уметь работать с Cassandra

Уметь обучать и применять ML модель на стримах

Параметры обработки ПОТОКОВЫХ данных



Mathias Verraes
@mathiasverraes

...

There are only two hard problems in distributed systems: 2. Exactly-once delivery 1. Guaranteed order of messages 2. Exactly-once delivery

9:40 PM · Aug 14, 2015 · Twitter for Android

7.2K Retweets 144 Quote Tweets 6.2K Likes

1. Задержка

- До секунды
- До нескольких часов

2. Консистентность

- Дубликаты
- Исторический порядок

3. Масштабируемость

- Распределенность
- Аналитика

Содержание курса

1. Spark Streaming. Тестовые стримы.
2. Kafka. Архитектура.
3. Spark Streaming. Чтение Kafka.
4. Spark Streaming. Sinks.
5. Spark Streaming. Stateful streams.
6. Lambda архитектура. Spark Streaming + Cassandra.
7. Spark ML. Аналитика признаков в пакетном режиме.
8. Spark Streaming + Spark ML + Cassandra.

Спасибо!
Каждый день
вы становитесь
лучше :)

