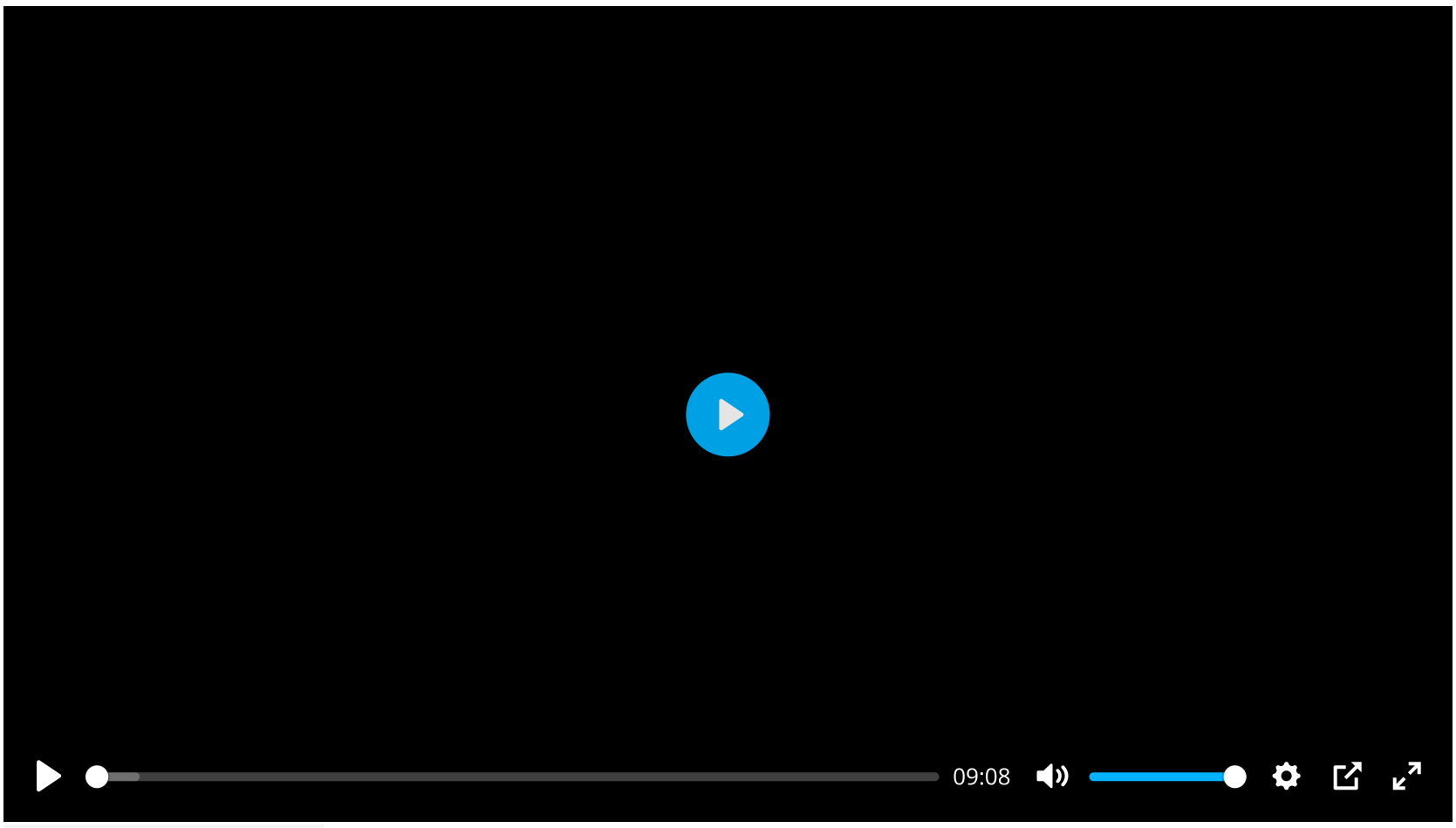
4.2 Что такое Spark и зачем он нужен DE. Введение в Resilient Distributed Datasets



Презентация к уроку 4.2

Текстовая расшифровка видео:

ЧТО ТАКОЕ SPARK И ЗАЧЕМ ОН НУЖЕН DE. ВВЕДЕНИЕ В RDD

План:

- Зачем нужен Spark;
- Что предлагает Spark.

Зачем нужен Spark

Зачем нужен Spark и почему нельзя взять Hadoop, слить данные в базу данных и т.д.? Для начала поговорим о проблемах с MapReduce.

Проблемы с MapReduce:

- Вместо написания бизнес-логики приходится концентрироваться на том, как ее натянуть на MapReduce.
- Все промежуточные результаты вычислений сохраняются на диск, который намного медленнее памяти.
- Даже самые базовые операции (loin, aggregate и т.д.) требуют написания руками довольно большого количества кода.

Что предлагает Spark?

- Проблема упрощения кода для **параллельных вычислений** актуальна как **в рамках одной машины,** так и **на гигантских кластерах**. Spark помогает решить данную проблему, так как дает общий фреймворк для вычислений на разных скейлах.
- Иногда данные поступают **потоком** и необходимо реагировать на них **как можно быстрее**. Современный Spark неплохо умеет это делать, его часто используют для потоковых вычислений.

- Современный Hive как механизм выполнения SQL-запросов использует те механизмы, которые предлагают Spark. Внутри себя он использует Spark, потому что в ядре технологии **Spark реализовали хороший оптимизатор, позволяющий удобно отлаживать и оптимизировать SQL-запросы**. Когда мы пишем базовые операции на данных в Spark, у нас проходят несколько стадий разных оптимизаций, и SQL-запросы выполняются гораздо эффективнее.
- Spark из «коробки» предлагает разные готовые компоненты. У него есть **расширенная стандартная библиотека** и поддержка дополнительных плагинов для **работы с графами** и **машинного обучения**.

Как вам урок? 😠 😐 🔥

Изучил, далее ›

Дедлайн 07 июля, 23:59 Мск