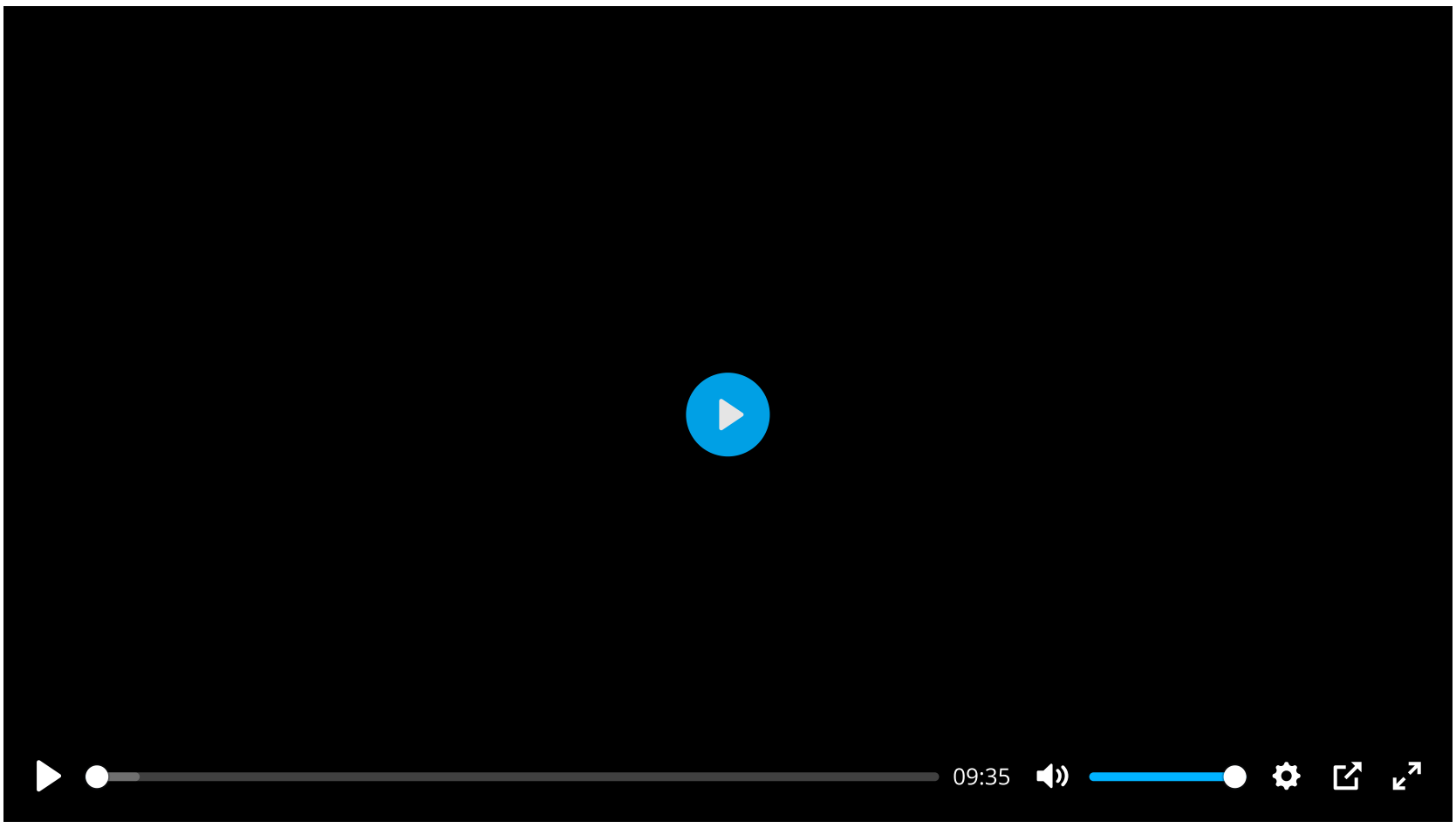
4.3 Dataframe'ы и преобразования



Текстовая расшифровка видео:

ПАРА СЛОВ О UDF

План:

- User defined functions.

User defined functions

UDF (User defined functions) существует и для **Python API,** и для **Scala API**, причем и там и там они считаются **антипаттерном**.

При применении UDF **отключатся все оптимизации**, предоставляемые, например, Catalyst.

Существует возможность писать UDF на голом Python. Однако она почти никогда не нужна в реальной жизни.

В современном Spark с датафреймами и PyArrow в случае крайней нужды можно использовать Pandas UDF.

Главная концепция, которая приносит при вычислении UDF, – **векторизация**. Если вам удастся максимально векторизовать код, то вполне возможно его ускорить в несколько раз, иногда даже в десятки.

ИТОГИ

Мы:

- Рассмотрели RDD и DataFrame, в частности их плюсы, минусы, подводные камни. **RDD** – низкоуровневая коллекция без особых оптимизаций; **DataFrame** – оптимизированная узкоспецифичная структура, которая во многих задачах проявляет бóльшую эффективность.
- Узнали о разнице между запуском JVM-байткода и Python.
- Научились читать данные и проводить базовые операции на DataFrame, чтобы их очистить и

- Рассмотрели несколько более сложных кейсов из практики и набросили на UDF.

Как вам урок?

Изучил, далее

Дедлайн 07 июля, 23:59 Мск