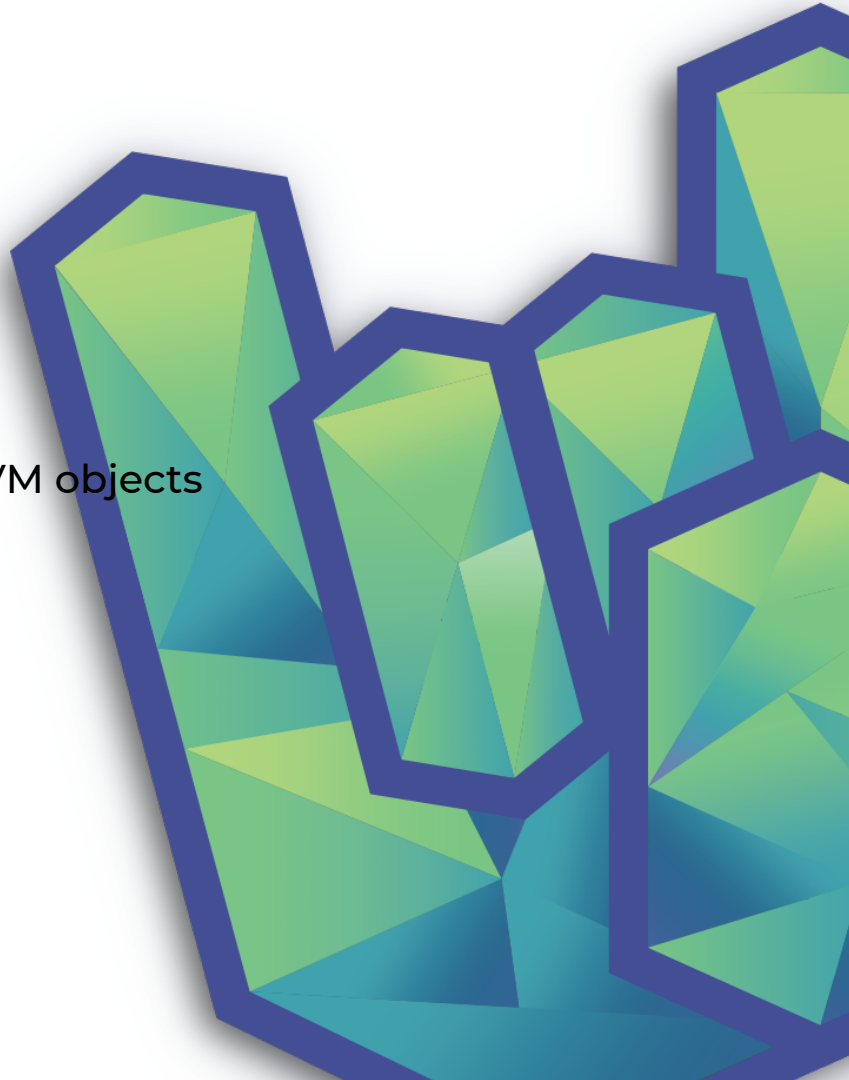
Datasets

Datasets, Part 2

Joining and grouping

Datasets

Typed DataFrames: distributed collection of JVM objects



Datasets

Distributed collection of JVM objects

Most useful when

- we want to maintain type information
- we want clean concise code
- our filters/transformations are hard to express in DF or SQL

Avoid when

- performance is critical: Spark can't optimize transformations

Spark rocks