# Pre-Partitioning

# Objective

Optimization: partition your data so that Spark doesn't have to

# To Remember

## Partition your data early so that Spark doesn't have to

- make the joined DFs share the same partitioner, e.g. partition by the same column
- decorate the joined DF later (especially if you have lots of transformations)

## Partitioning late is bad

- at best: same perf as Spark out-of-the-box
- at worst: worse performance than not partitioning at all

# Spark rocks