Bucketing

Objective

Split data intelligently before a join

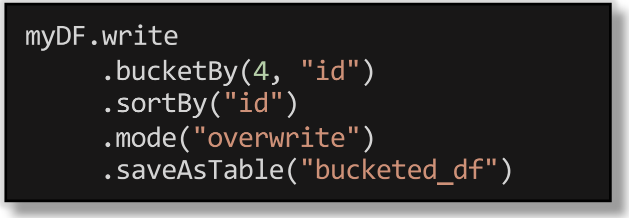
Benefit on multiple joins/groups

Bonus: bucket pruning

To Remember

Split the data by the columns you will later use for joins

- n buckets per partition



```
myDF.write
     .bucketBy(4, "id")
     .sortBy("id")
     .mode("overwrite")
     .saveAsTable("bucketed_df")
```

Almost as expensive as a regular shuffle

Subsequent joins will be much faster and will not incur shuffles

Same benefit for groupBy

- pre-bucket your DFs by the columns you will use for grouping
- subsequent groupings will be much faster

Bonus: bucket pruning

- Spark will only search a subset of your DF when you filter by the column you bucketed by

Spark rocks