# Fixing Data Skews

# Objective

Identifying straggling tasks

Diagnosing non-uniform data distributions

Applying the salting technique for uniform task distribution

# Fixing Data Skews

## Non-uniform data distribution => non-uniform task distribution

- straggling tasks: will delay the completion of a stage
- not solvable with more resources

## Salting

- explode the smaller RDD or DF with salt values for every row
- in the other RDD/DF, add a random value from the salt interval for every row
- join on the combined key = original key + salt column
- combine partial results if you need to

## How salting works

- data is now distributed by n+1 keys, one of which is uniform
- the larger the salt interval, the less skewed tasks
- the larger the salt interval, the larger the shuffle size

## Also works for groups

# Spark rocks