

**I2I**

# **Transformations**



# Objective

Iterator-to-iterator transformations

Manipulate RDD partitions in an arbitrary fashion



# To Remember

I2I transformation: manipulate each partition with an iterator function

## Benefits

- any transformation per-partition is a narrow transformation
- if partitions are too large, Spark will spill to disk what can't fit

## Anti-patterns

- collecting the iterator in memory
- multiple passes over the data

## Lessons

- traverse the iterator ONCE
- don't use collection conversions

**Spark rocks**

