

RDD By-Key Functions



Objective

For key-value RDDs

Overview of by-key functions

Performance implications (and more)



To Remember

Not all by-key functions are equal

`groupByKey` is the most intuitive, but the most dangerous

- shuffles all the data
- can cause long ("straggler") tasks and memory issues/OOMs in case of data skews

`reduceByKey`

- like the standard collections API
- faster and safer: reduces locally on executor first, shuffles less data

`foldByKey` – similar perf

`aggregateByKey` – similar perf, more powerful API

`combineByKey`

- the most general combiner function
- if used correctly, the most efficient + more efficient in subsequent operations e.g. joins
- as (or more) dangerous as `groupByKey` if functions are not "reductions"
- all other by-key functions implemented using this

Spark rocks

