RDD Cogroups

Objective

Use cogroup to speed up joins

To Remember: Cogroup

Makes sure ALL the RDDs share the same partitioner

Particularly useful for multi-way joins

- all RDDs are shuffled at most once
- RDDs are never shuffled again if cogrouped RDD is reused

Keeps the entire data - equivalent to a full outer join

- for each key, an iterator of values is given
- if there is no value for a "column", the respective iterator is empty

Spark rocks