RDD Abstraction

- immutable, partitioned collection of [[Key-Value Pairs]]
- fault tolerance via lineage-based re-computation

Operations

- coarse-grained deterministic operations
  - transformations
  - actions

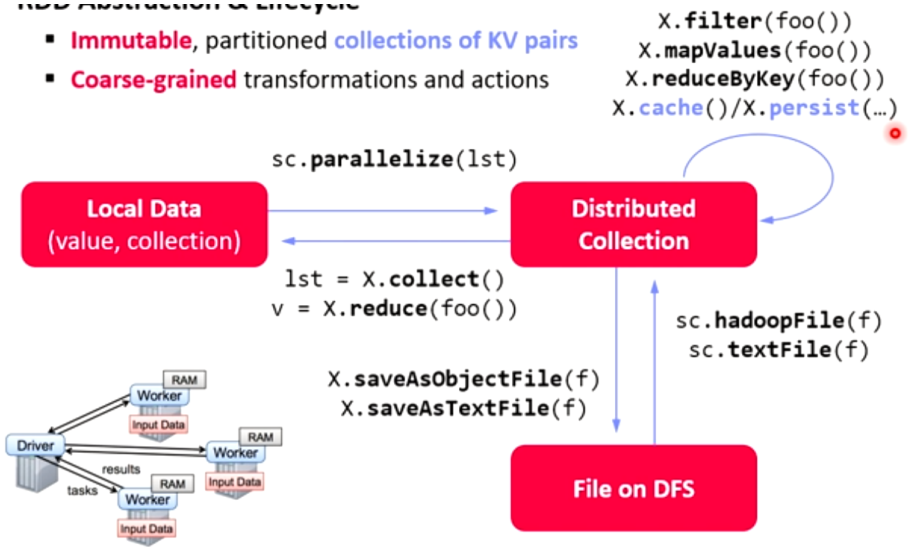
Type	Examples
Transformation (lazy)	map, hadoopFile, textFile, flatMap, filter, sample, join, groupByKey, cogroup, reduceByKey, cross, sortByKey, mapValues
Action	reduce, save, collect, count, lookupKey

-

Distributed Caching

- fraction of worker memory used for caching
- different storage levels
  - (e.g., mem/disk x serialization x compression)

RDD Lifecycle



-