

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)	<code>map</code> , <code>hadoopFile</code> , <code>textFile</code> , <code>flatMap</code> , <code>filter</code> , <code>sample</code> , <code>join</code> , <code>groupByKey</code> , <code>cogroup</code> , <code>reduceByKey</code> , <code>cross</code> , <code>sortByKey</code> , <code>mapValues</code>
Action	<code>reduce</code> , <code>save</code> , <code>collect</code> , <code>count</code> , <code>lookupKey</code>

Distributed Caching

- fraction of worker memory used for caching
- different storage levels

(e.g., mem/disk x serialization x compression)

RDD Lifecycle

RDD Abstraction & Lifecycle

- **Immutable**, partitioned **collections of KV pairs**
- **Coarse-grained** transformations and actions

```
X.filter(foo())  
X.mapValues(foo())  
X.reduceByKey(foo())  
X.cache()/X.persist(...)
```

