Solution: Resilient Distributed Datasets (RDDs)

- RDDs
  - Immutable, partitioned collection of records
  - Built through coarse grained, ordered transformations (map, join …)

- Fault Recovery?
  - Lineage!
    - Log the coarse grained operation applied to a partitioned dataset
    - Simply recompute the lost partition if failure occurs!
    - No cost if no failure