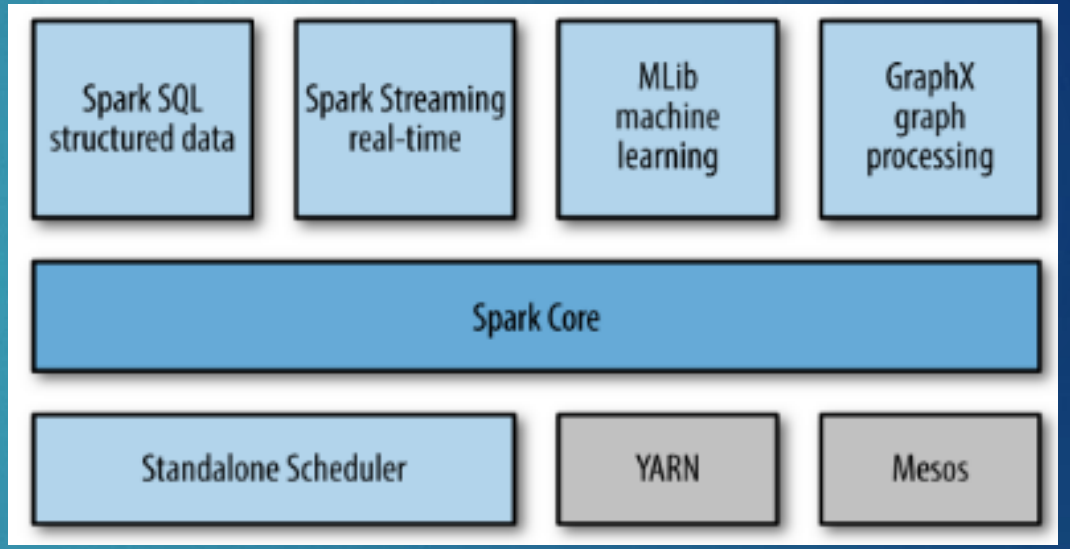
# Apache Spark

KYU CHO

11/30/16

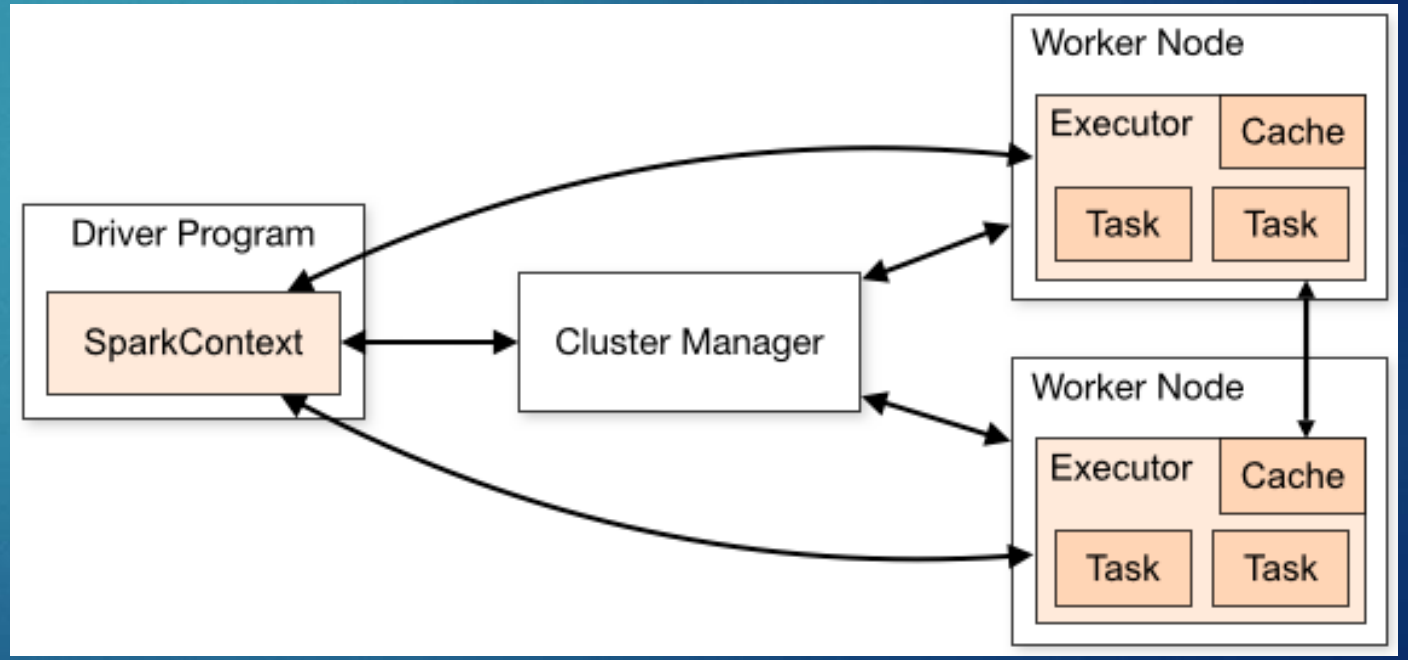
# Environment

- Spark Streaming
  - Analyze real data from the web
- SparkSQL
  - hive contexts
  - deal with structure data
  - SQL data on top of it
  - data warehouse
- MLLib
  - machine learning algorithms
- GraphX
  - managing networks
    - ex) social networks



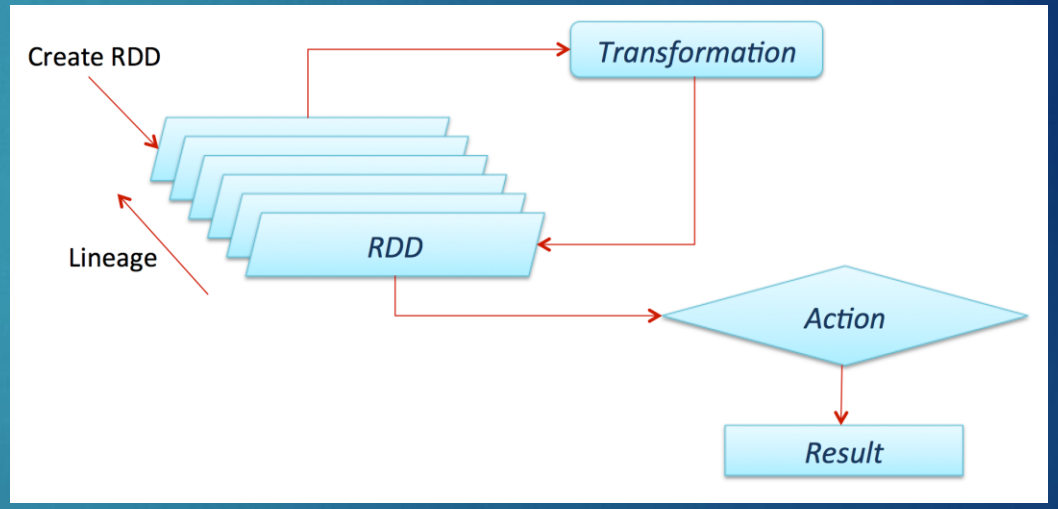
# Advantages

- Cluster Manager -> fault tolerant automatically
- Faster than Hadoop MapReduce
- Memory management

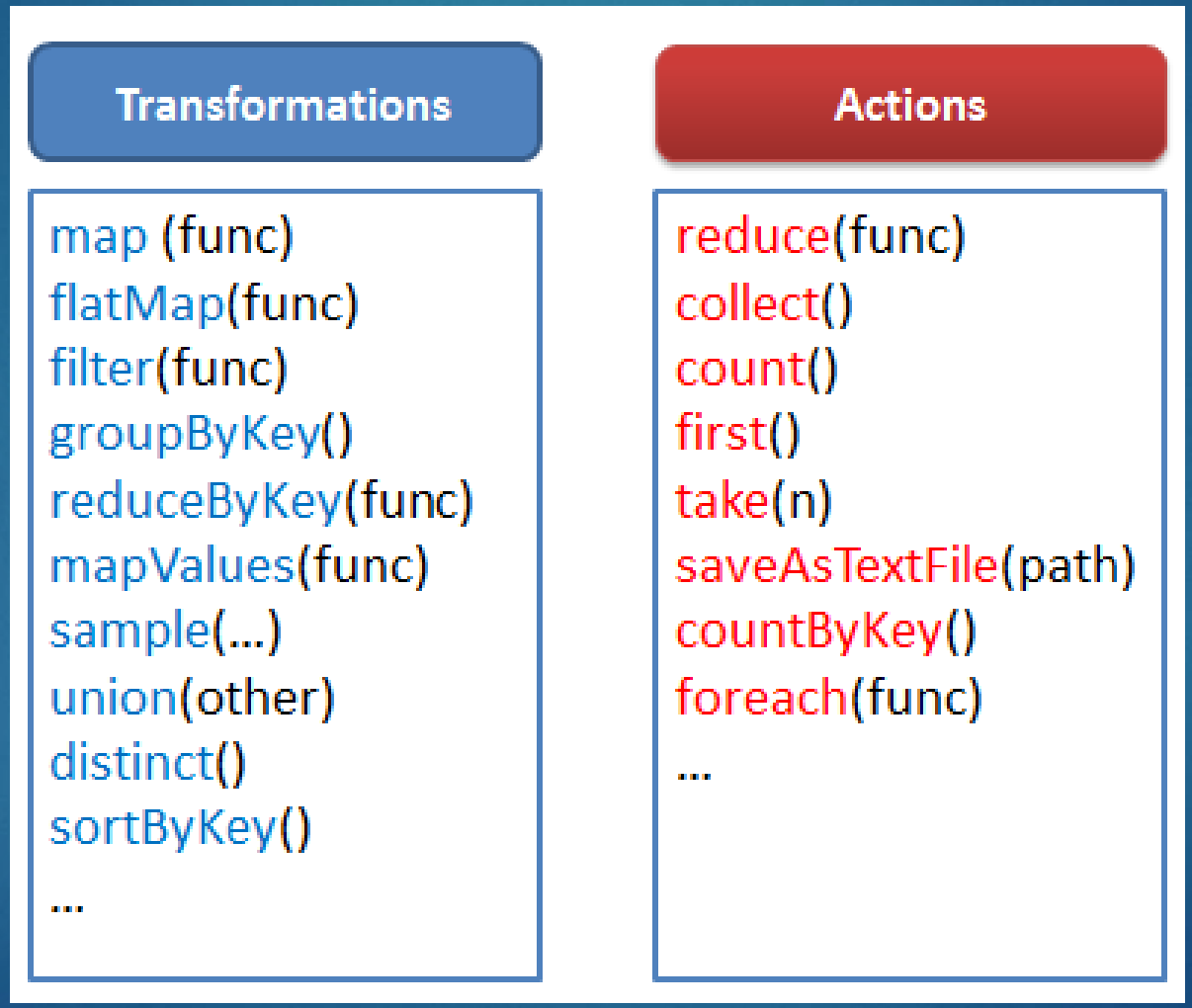


# RDD

- DAG Engine (directed acyclic graph) optimizes workflows
  - wait till user ask to deliver the result
  - figure out the optimal path
- RDD (Resilient Distributed Dataset)
  - data transformation-> quick and efficient



# Transformation and Action



# Demo Time

- Thanks