Spark Cheat Sheet

cloudera®

SC Operations

Load Files into RDDs

sc.textFile
sc.wholeTextFiles

RDD Operations

Transformations

rdd.map()
rdd.flatMap()
rdd.filter()
rdd.join()
rdd.groupBy()
rdd.reduceByKey()

Actions

rdd.count()
rdd.take()
rdd.collect()
rdd.saveAsTextFile()

Persistance

rdd.cache()
rdd.persist()
rdd.checkpoint()

Debugging

rdd.toDebugString()

Streaming

Setup

ssc = StreamingContext(conf, interval)
ssc.socketTextStream()
ssc.start()
ssc.awaitTermination()

Submitting Jobs

Scala/Java

$ spark-submit —class myclass.class myjar.jar [params]

Python

$ spark-submit myfile.py [params]

Shared Variables

Broadcast Variables

sc.broadcast()
.value()

Accumulators

sc.accumulator()
.add()