**Transform, Stage, and Store**

* **Load RDD data from HDFS for use in spark applications:**

text\_file = sc.textFile("/home/cloudera/hdfs\_data")

--Input HDFS directory

counts = text\_file

.flatMap(lambda line: line.split(" "))

-- splits the lines by space

.map(lambda word: (word, 1))

-- maps the word(KEY) with 1(VALUE)

.reduceByKey(lambda a, b: a + b)

-- aggregates it based on the KEY

counts.saveAsTextFile("/home/cloudera/hdfs\_output")

-- Output HDFS directory