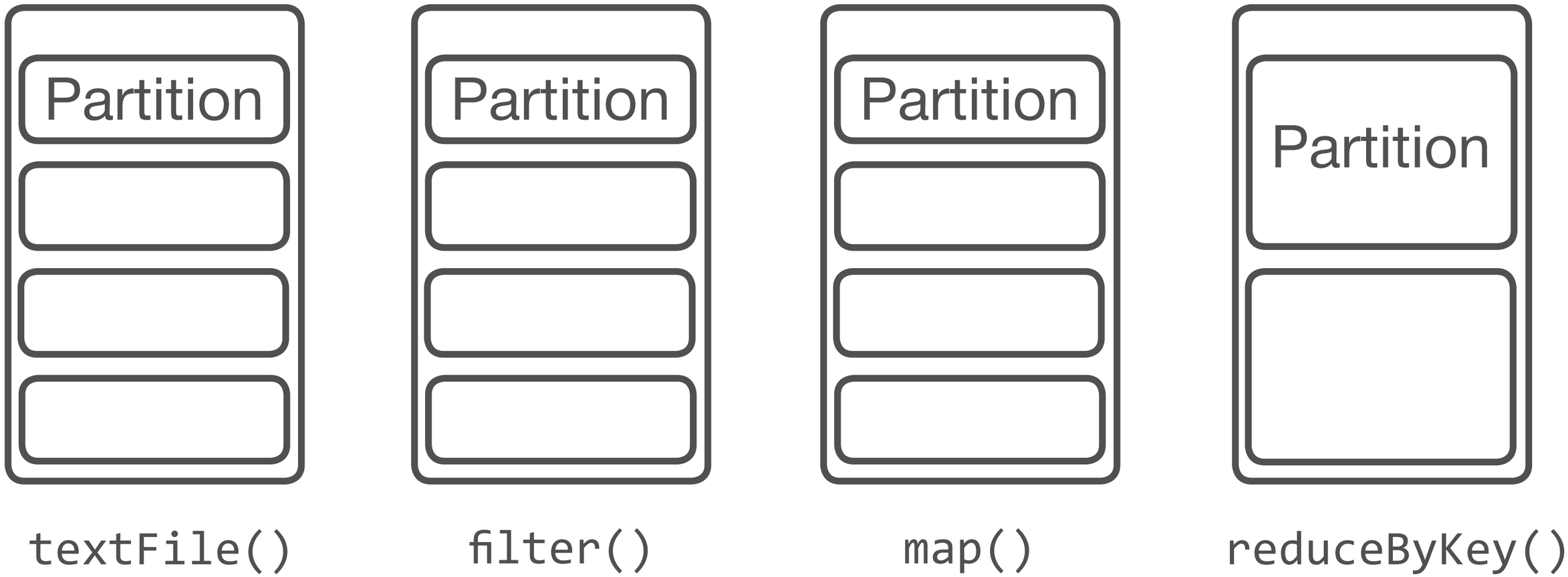
Spark Execution Model

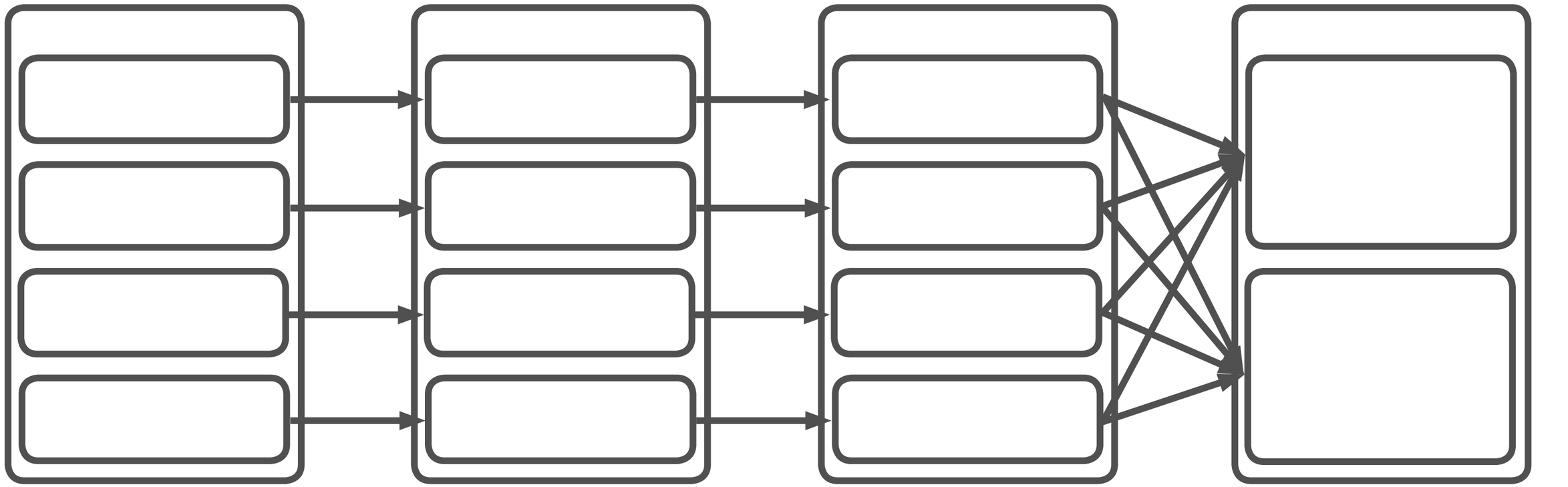
192.168.0.38	WARNING	Something bad could happen
192.168.0.88	INFO	Just an info message passing by
192.168.0.5	WARNING	Something bad could happen
192.168.0.36	ERROR	When production fails in despair, whom are you gonna call?
192.168.0.27	INFO	Just an info message passing by

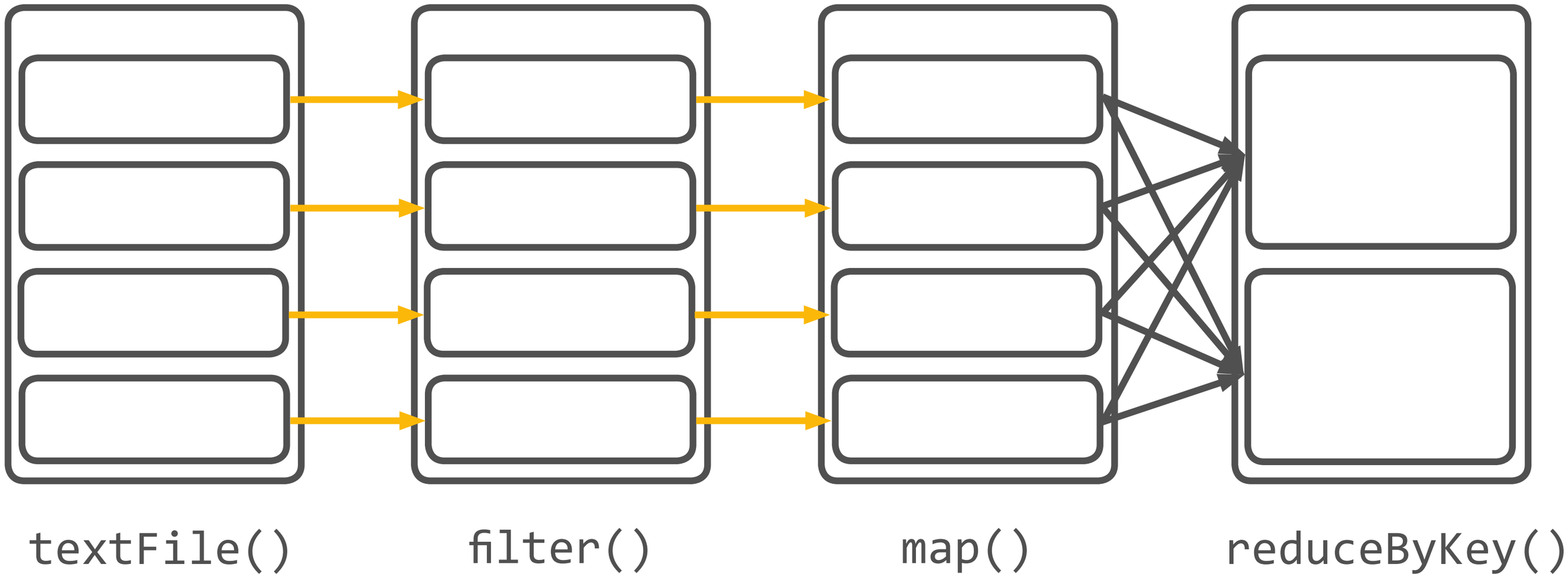
192.168.0.38	USA
192.168.0.88	RUSSIA
192.168.0.5	CHINA
192.168.0.36	USA
192.168.0.27	RUSSIA

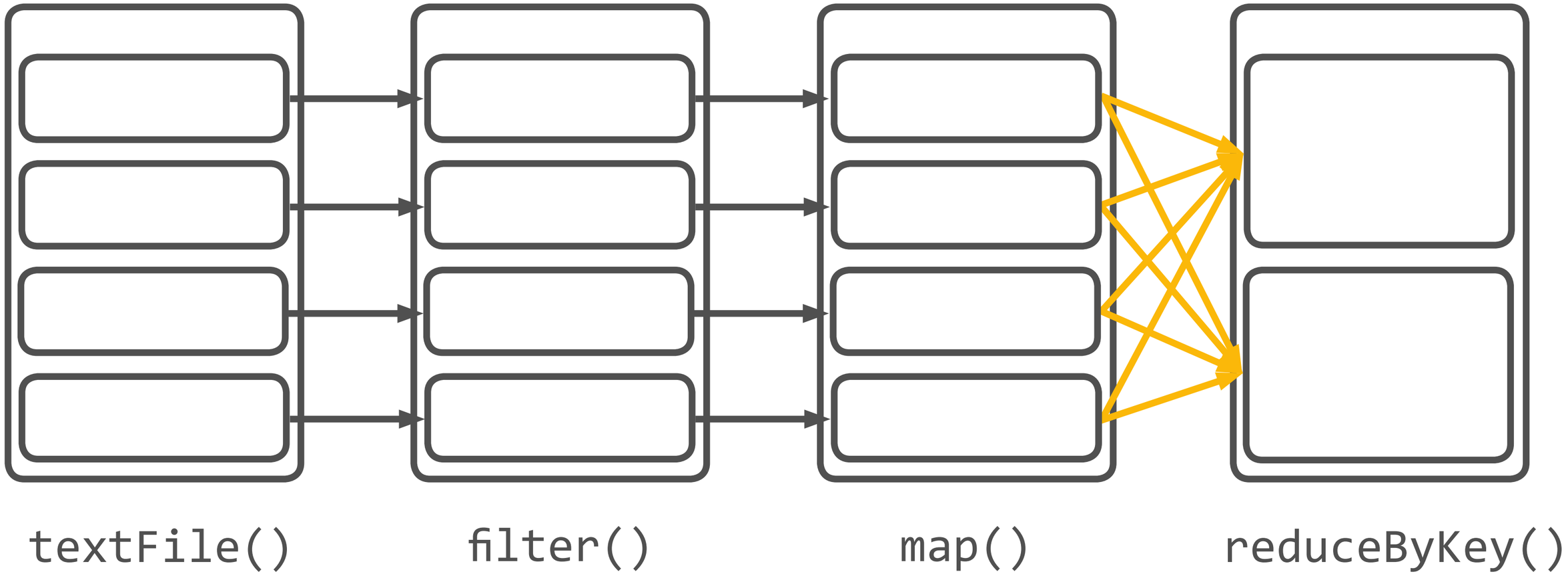
```
logs = sc.textFile("log.txt")
        .filter(lambda x: "INFO" not in x)
        .map(lambda x: (x.split("\t")[1], 1))
        .reduceByKey(lambda x, y: x + y)

logs.collect()
```

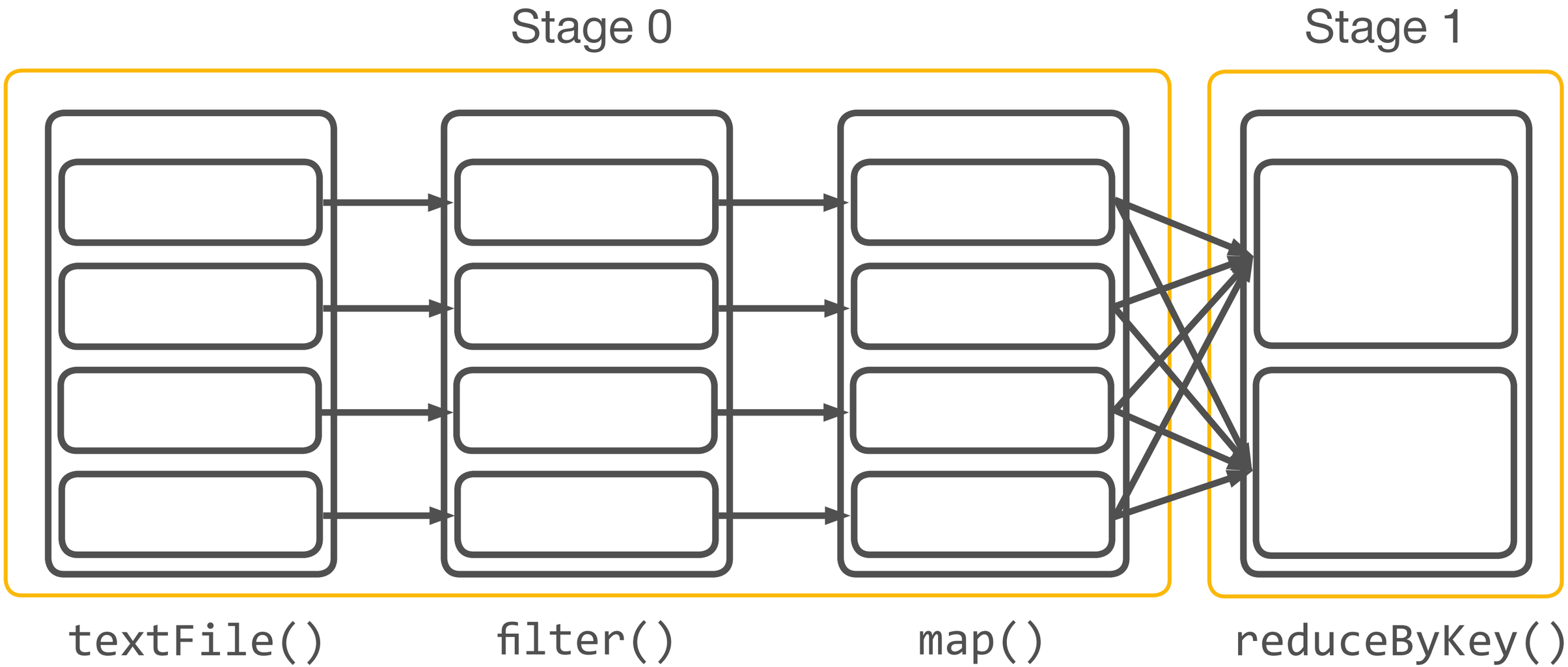

RDD
RDD
RDD
RDD
textFile()
filter()
map()
reduceByKey()


Partition
Partition
Partition
Partition
textFile()
filter()
map()
reduceByKey()


textFile()
filter()
map()
reduceByKey()


textFile()
filter()
map()
reduceByKey()


textFile()
filter()
map()
reduceByKey()


Stage 0
Stage 1
textFile()
filter()
map()
reduceByKey()

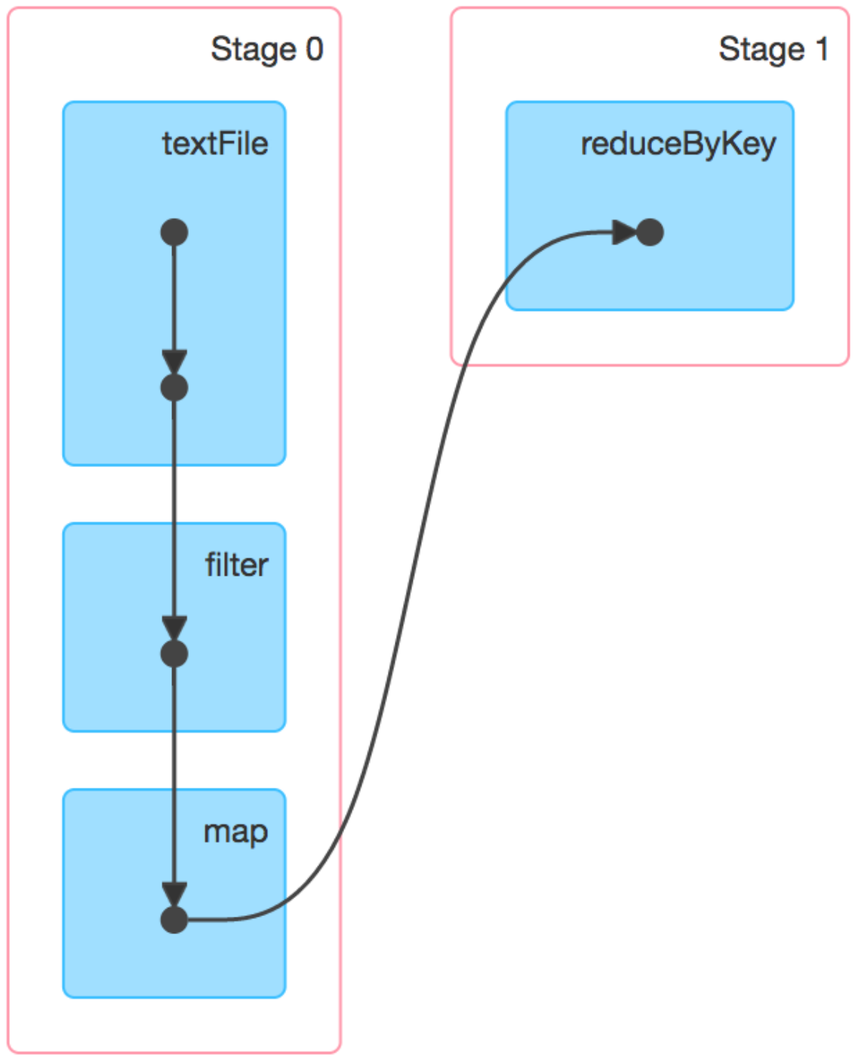
Details for Job 0

Status: SUCCEEDED

Completed Stages: 2

▸ Event Timeline

▾ DAG Visualization



Narrow transformations	Wide transformations
map	cogroup
mapValues	groupWith
flatMap	join
filter	leftOuterJoin
mapPartitions	rightOuterJoin
mapPartitionsWithIndex	groupByKey
	reduceByKey
	combineByKey
	distinct
	intersection
	repartition
	coalesce

Summary

- RDD consists of partitions
- Partition is an atomic unit of parallelism
- RDD may have narrow or wide dependencies
- Narrow dependencies can be pipelined
- Wide dependencies cause shuffles
- Shuffles are expensive