Apache Spark 101

Nirmal Ranganathan

General purpose engine for

large-scale data processing

Brief History

- 2009 - AMPLab at the University of California, Berkeley
- 2013 - Apache incubation
- 2014 - Top Level Project
- 600+ contributors

Interactive Queries

Stream Processing

Machine Learning

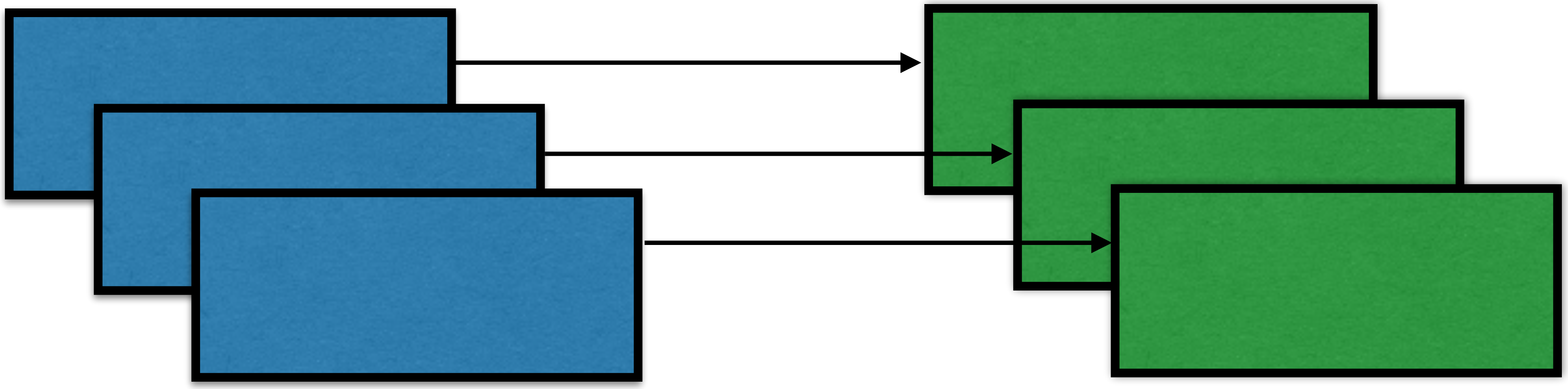
Why so popular?

Resilient Distributed Datasets (RDDs)

RDDs

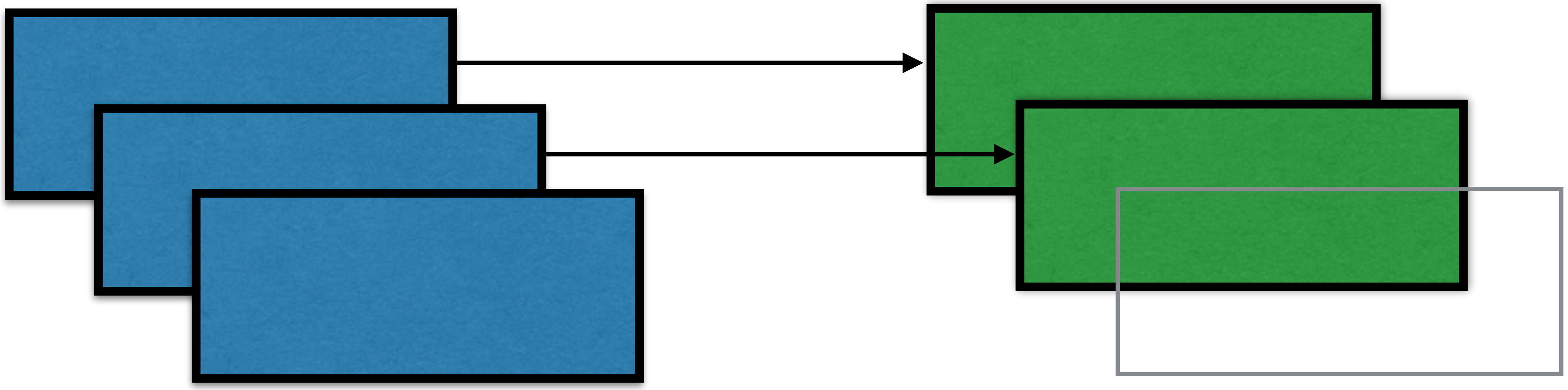
- **Transformations**
  - map
  - filter
  - groupByKey
  - ReduceByKey
- **Actions**
  - collect
  - reduce
  - count
  - save

Transformation



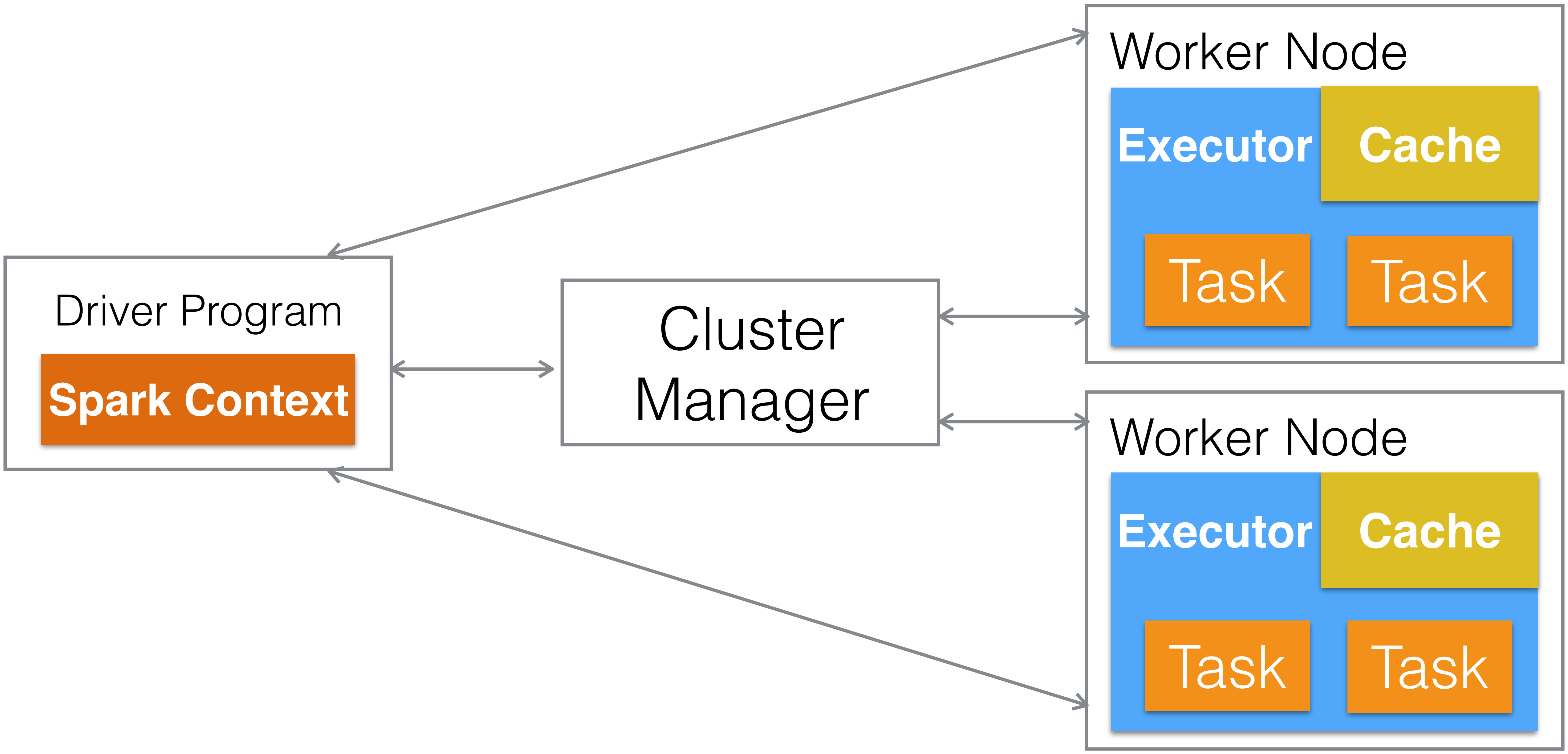
`map(func)`

Transformation

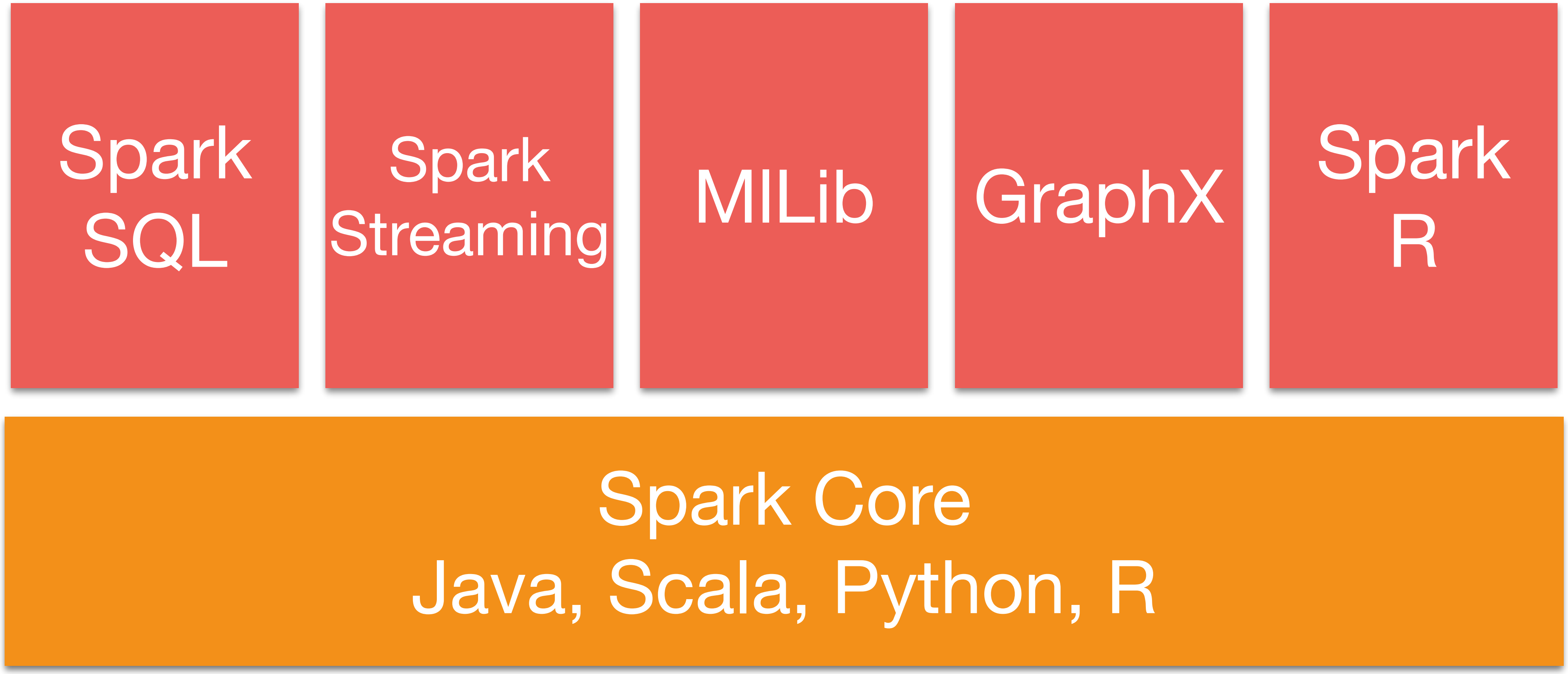


`filter(func)`

Components


Worker Node
Executor
Cache
Task
Task
Driver Program
Spark Context
Cluster Manager
Worker Node
Executor
Cache
Task
Task

Components



Spark SQL

Spark Streaming

MlLib

GraphX

Spark R

Spark Core
Java, Scala, Python, R

Word Count

```
text_file = sc.textFile(“source.txt”)

counts = text_file.flatMap(lambda line: line.split(" “)) \
                  .map(lambda word: (word, 1)) \
                  .reduceByKey(lambda a, b: a + b)

counts.saveAsTextFile(“hdfs://host:port/results”)
```

Word Count

```
val text_file = sc.textFile(“source.txt”)

val counts = text_file.flatMap(line => line.split(" "))
                 .map(word => (word, 1))
                 .reduceByKey(_ + _)

counts.saveAsTextFile(“hdfs://host:port/results”)
```