Introduction to Spark

Shannon Quinn

(with thanks to Paco Nathan and Databricks)

Quick Demo

we’ll run Spark’s interactive shell…

```
./bin/spark-shell
```

then from the “scala>” REPL prompt, let’s create some data…

```
val data = 1 to 10000
```

Quick Demo

create an **RDD** based on that data…

```
val distData = sc.parallelize(data)
```

then use a filter to select values less than 10…

```
distData.filter(_ < 10).collect()
```

API Hooks

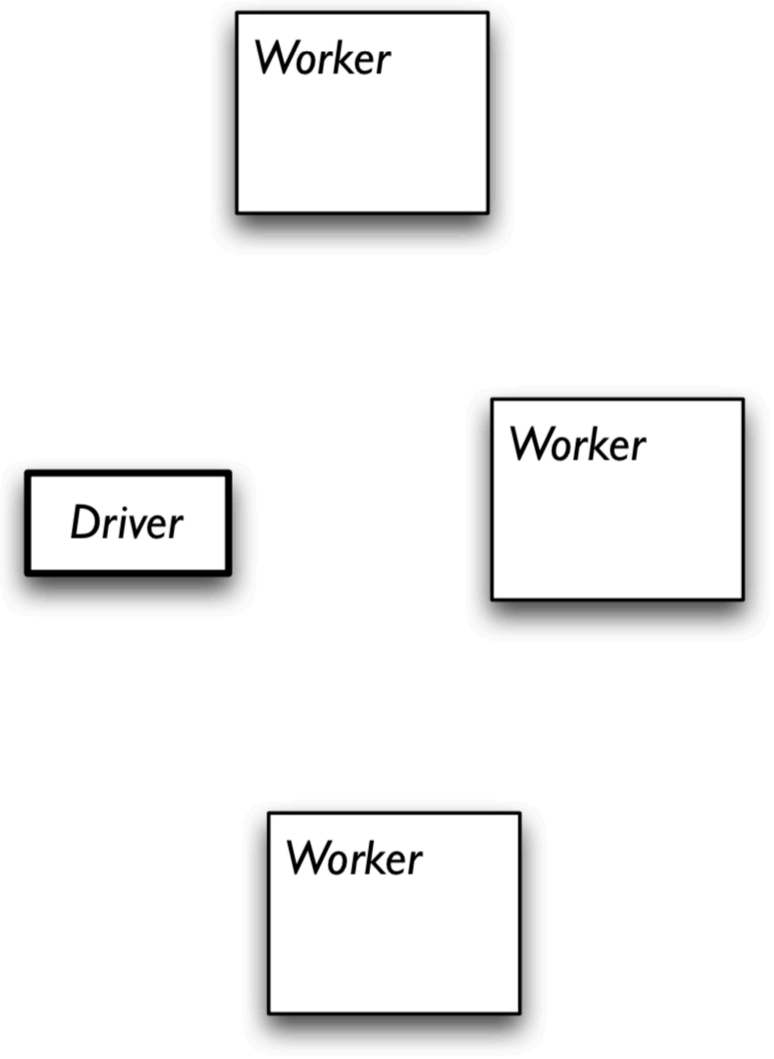
- Scala / Java
  - All Java libraries
  - *.jar
  - http://www.scala-lang.org
- Python
  - Anaconda: https://www.anaconda.com/download/
- ...R?
  - If you really want to
  - http://spark.apache.org/docs/latest/sparkr.html

Introduction

```
// load error messages from a log into memory
// then interactively search for various patterns
// https://gist.github.com/ceteri/8ae5b9509a08c08a1132

// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

Spark Structure

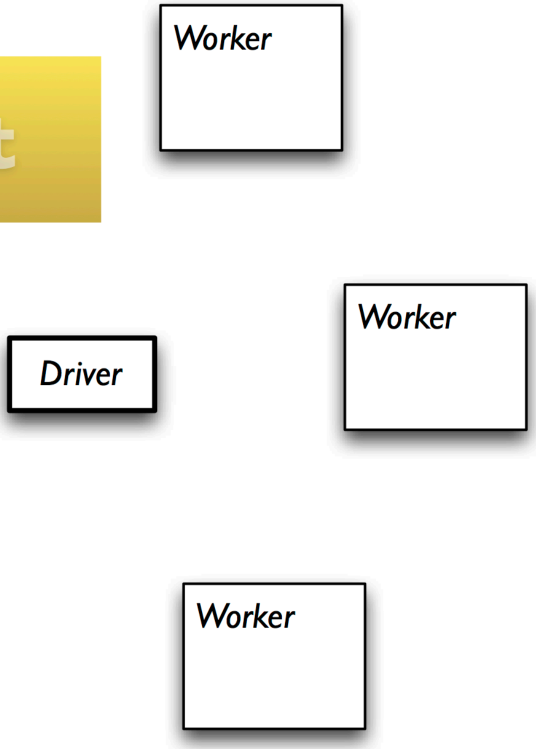
- Start Spark on a cluster
- Submit code to be run on it



Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

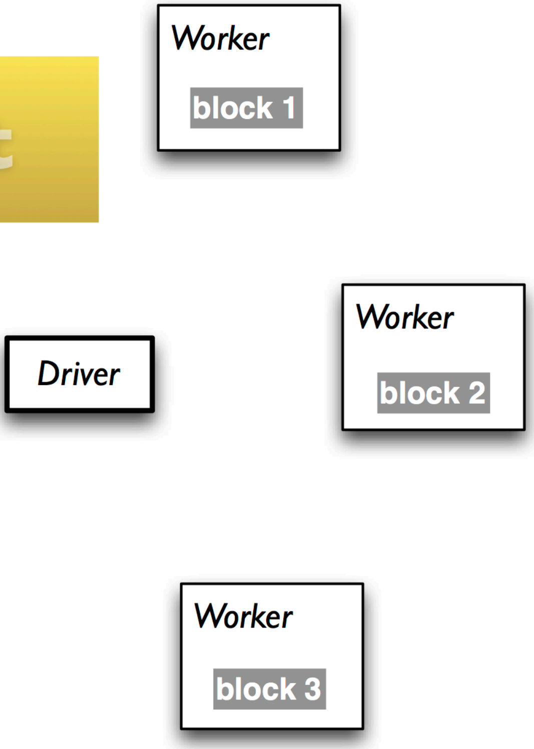
discussing the other part



Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

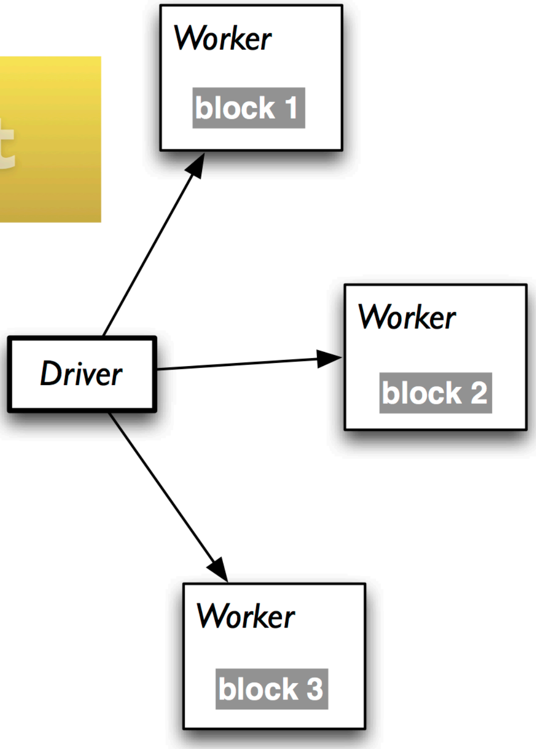
discussing the other part



Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

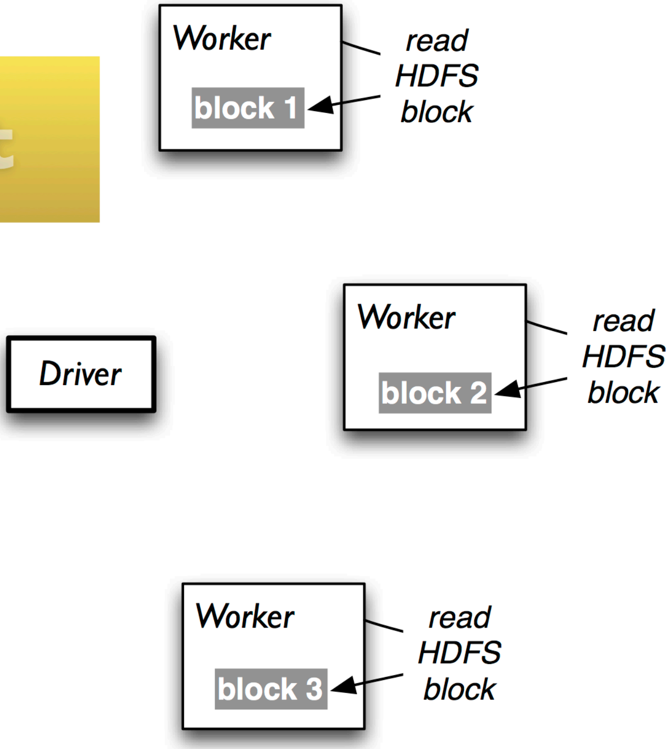
discussing the other part



Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

discussing the other part



Worker
block 1
read HDFS block

Driver

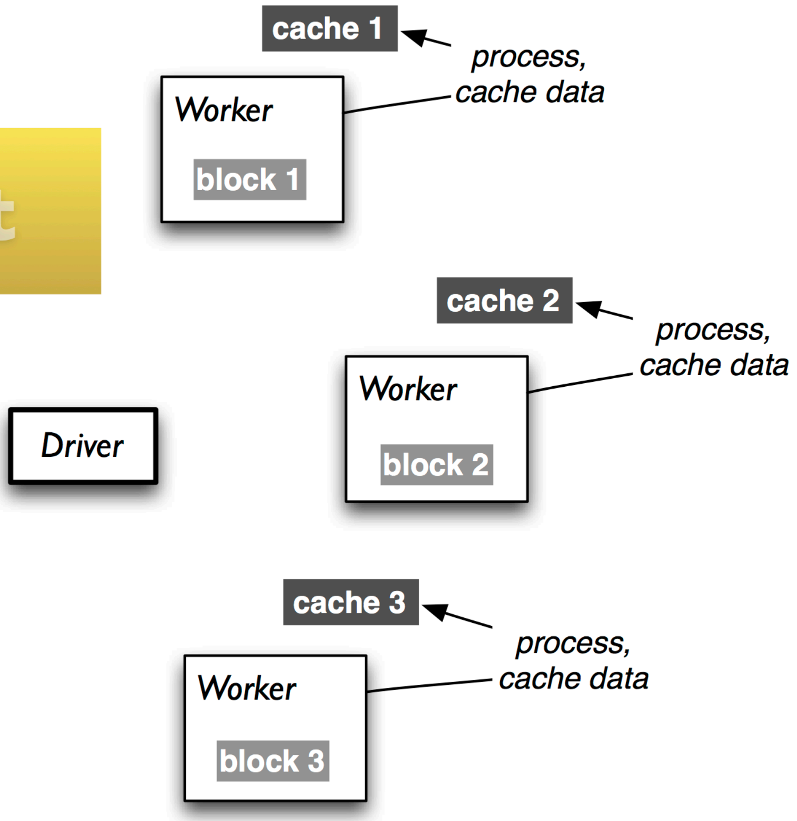
Worker
block 2
read HDFS block

Worker
block 3
read HDFS block

Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

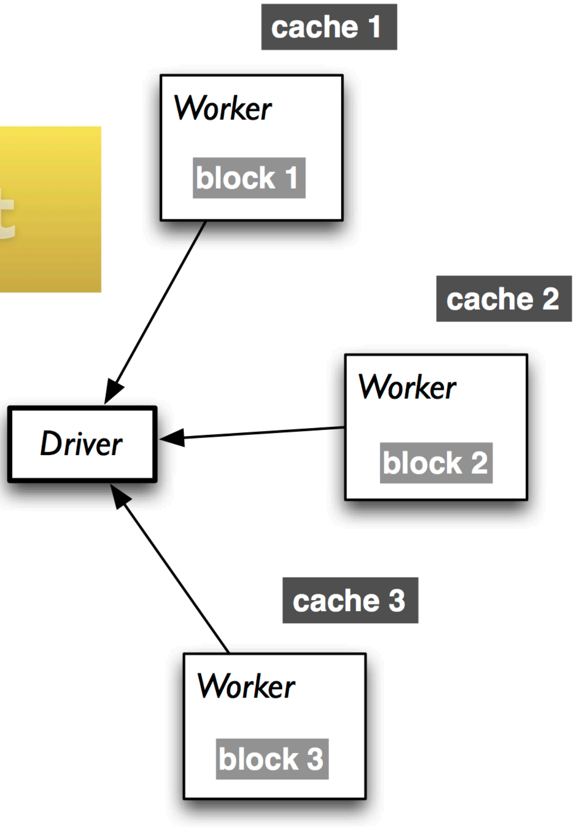
discussing the other part



Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```

discussing the other part

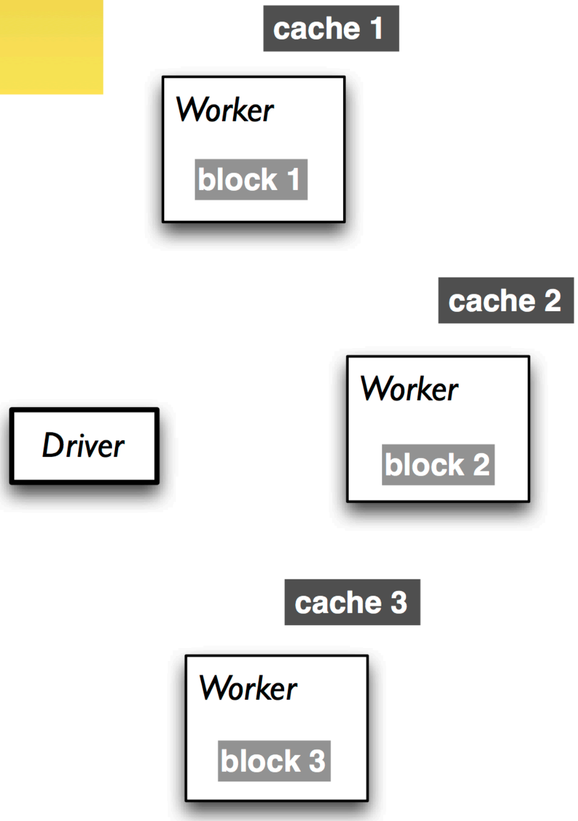


Spark Deconstructed: *Log Mining Example*

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()
```

discussing the other part

```
// action 2
messages.filter(_.contains("php")).count()
```

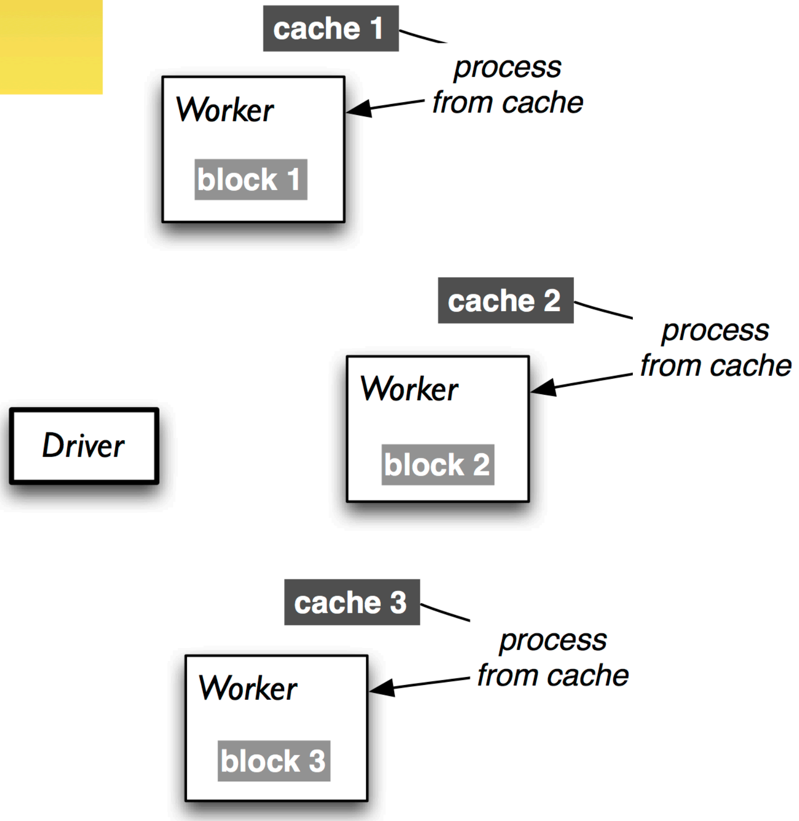


Spark Deconstructed: *Log Mining Example*

discussing the other part

```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains("mysql")).count()
```

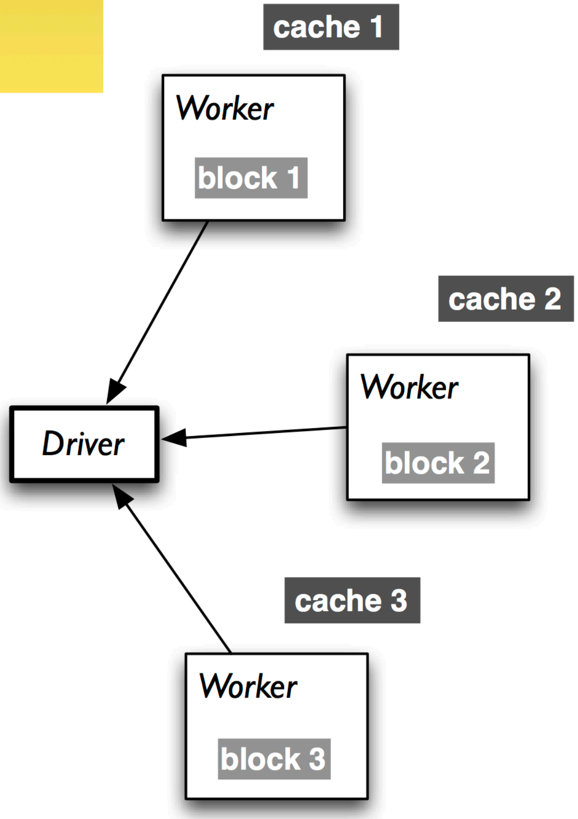
```
// action 2
messages.filter(_.contains("php")).count()
```



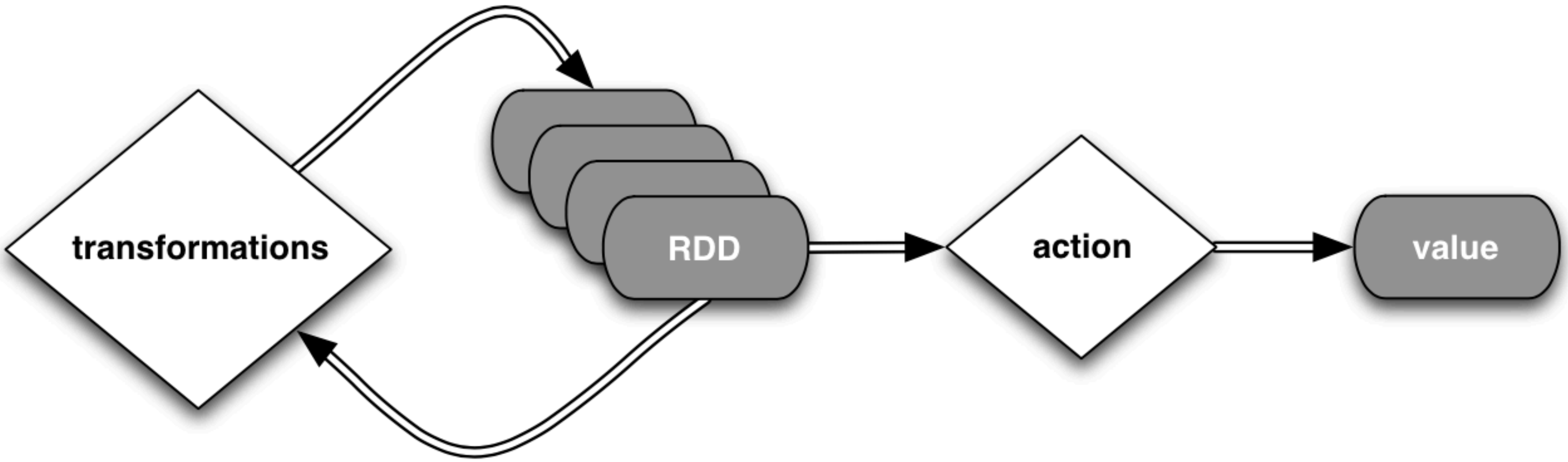
Spark Deconstructed: *Log Mining Example*

discussing the other part

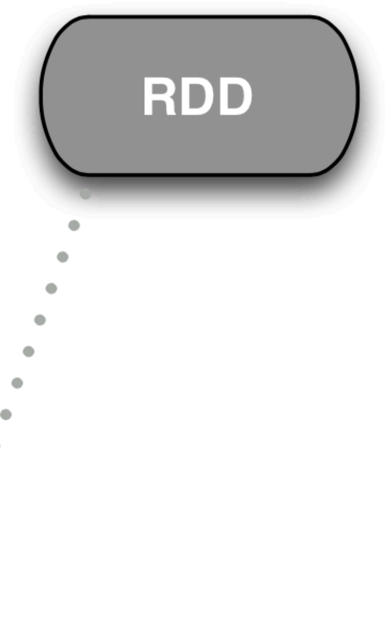
```
// base RDD
val lines = sc.textFile("hdfs://...")

// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()

// action 1
messages.filter(_.contains(“mysql")).count()

// action 2
messages.filter(_.contains("php")).count()
```



Another Perspective

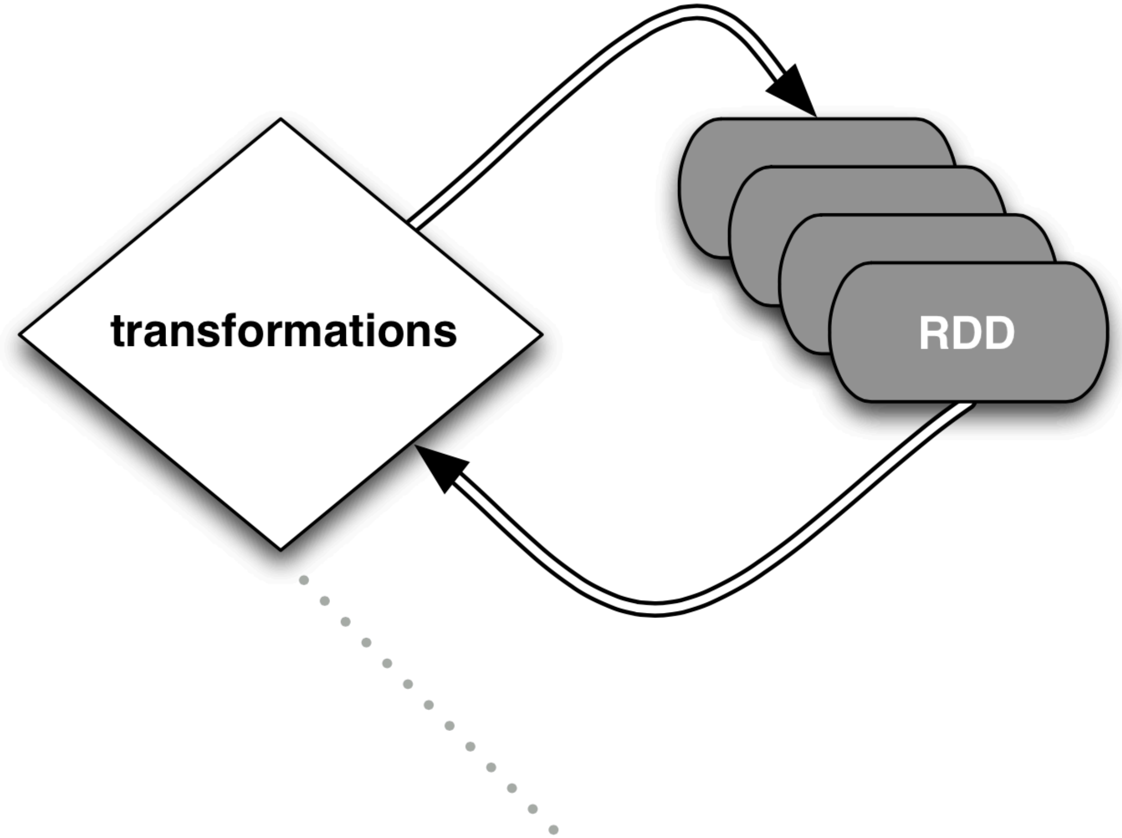


Step by step



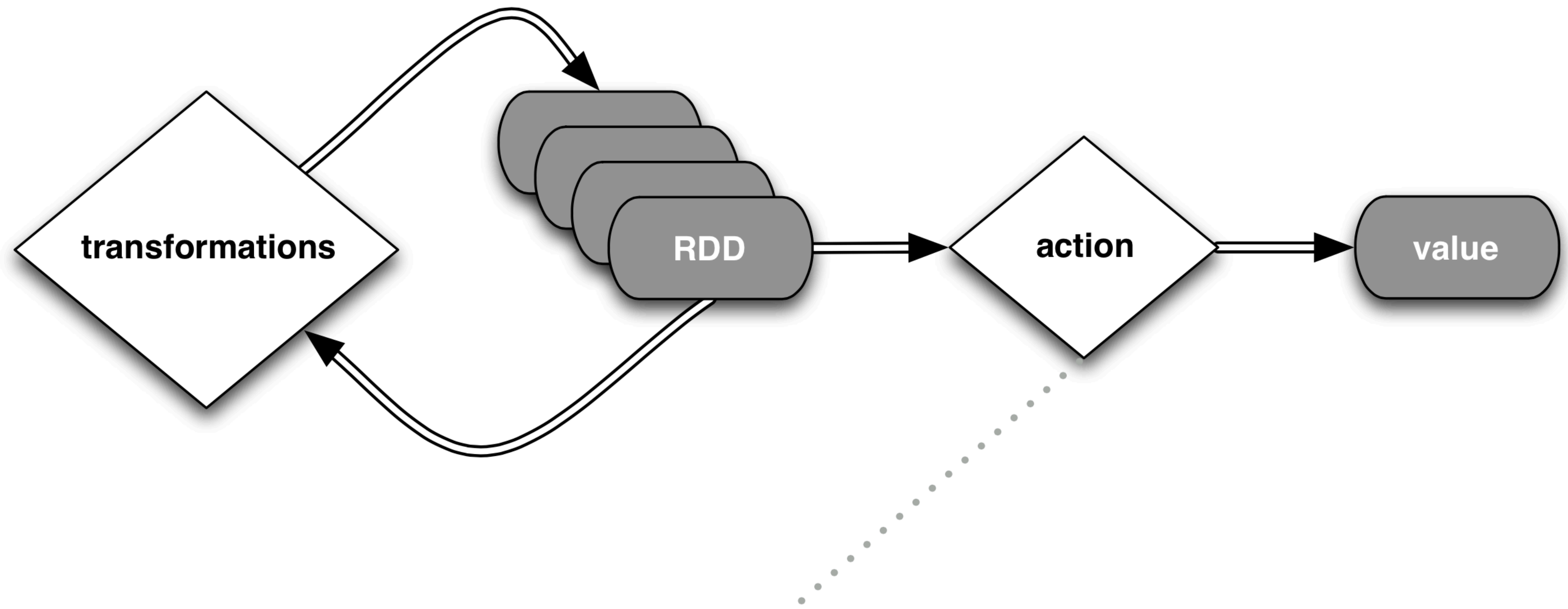
```
// base RDD
val lines = sc.textFile("hdfs://...")
```

Step by step



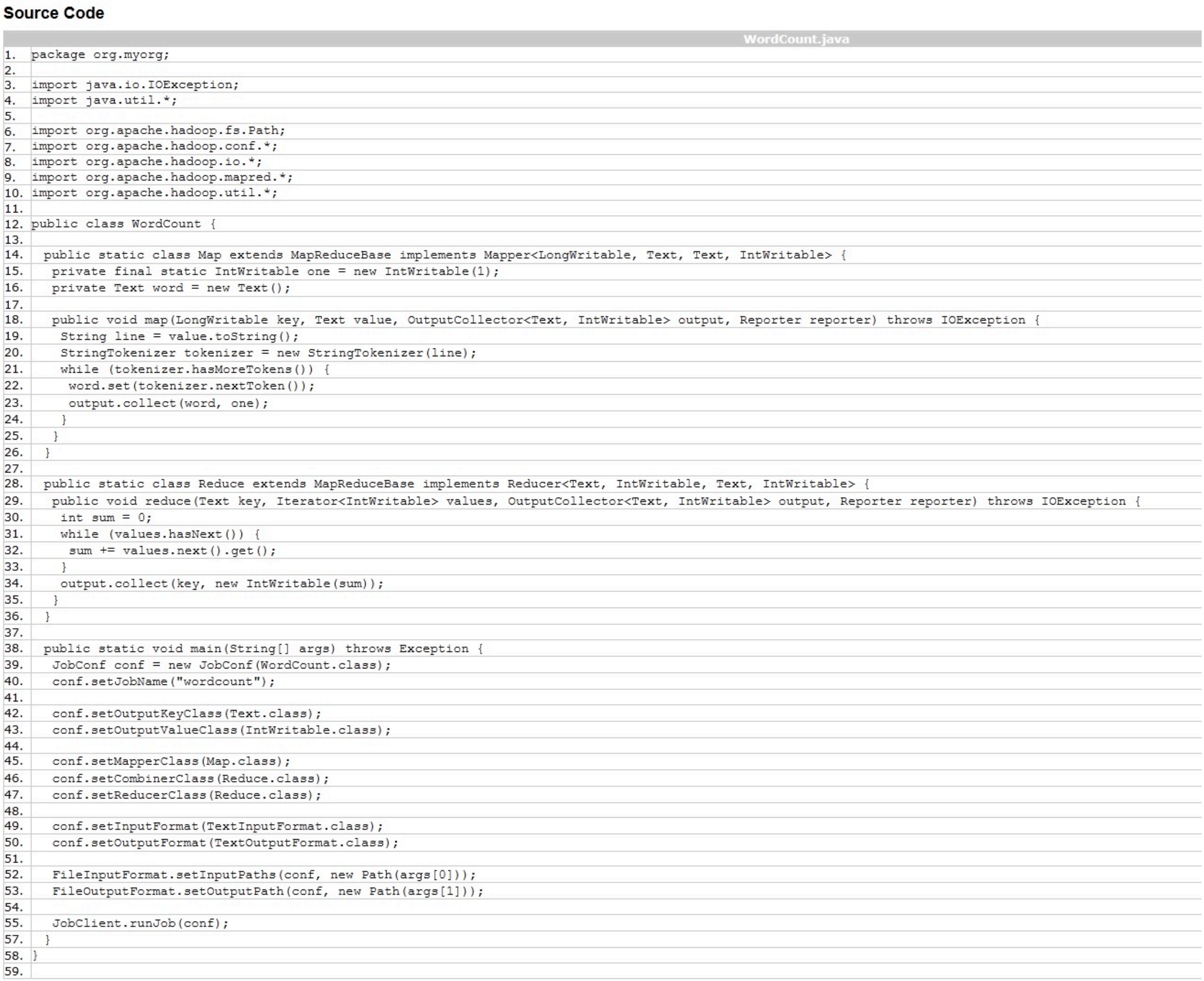
```
// transformed RDDs
val errors = lines.filter(_.startsWith("ERROR"))
val messages = errors.map(_.split("\t")).map(r => r(1))
messages.cache()
```

Step by step



```
// action 1
messages.filter(_.contains("mysql")).count()
```

Example: WordCount



Source Code

WordCount.java

```java
package org.myorg;

import java.io.IOException;
import java.util.*;

import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;

public class WordCount {

  public static class Map extends MapReduceBase implements Mapper<LongWritable, Text, Text, IntWritable> {
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
      String line = value.toString();
      StringTokenizer tokenizer = new StringTokenizer(line);
      while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        output.collect(word, one);
      }
    }
  }

  public static class Reduce extends MapReduceBase implements Reducer<Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text, IntWritable> output, Reporter reporter) throws IOException {
      int sum = 0;
      while (values.hasNext()) {
        sum += values.next().get();
      }
      output.collect(key, new IntWritable(sum));
    }
  }

  public static void main(String[] args) throws Exception {
    JobConf conf = new JobConf(WordCount.class);
    conf.setJobName("wordcount");

    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);

    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);

    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);

    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));

    JobClient.runJob(conf);
  }
}

```

Example: WordCount

Scala:

```scala
val f = sc.textFile("README.md")
val wc = f.flatMap(l => l.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
wc.saveAsTextFile("wc_out.txt")
```

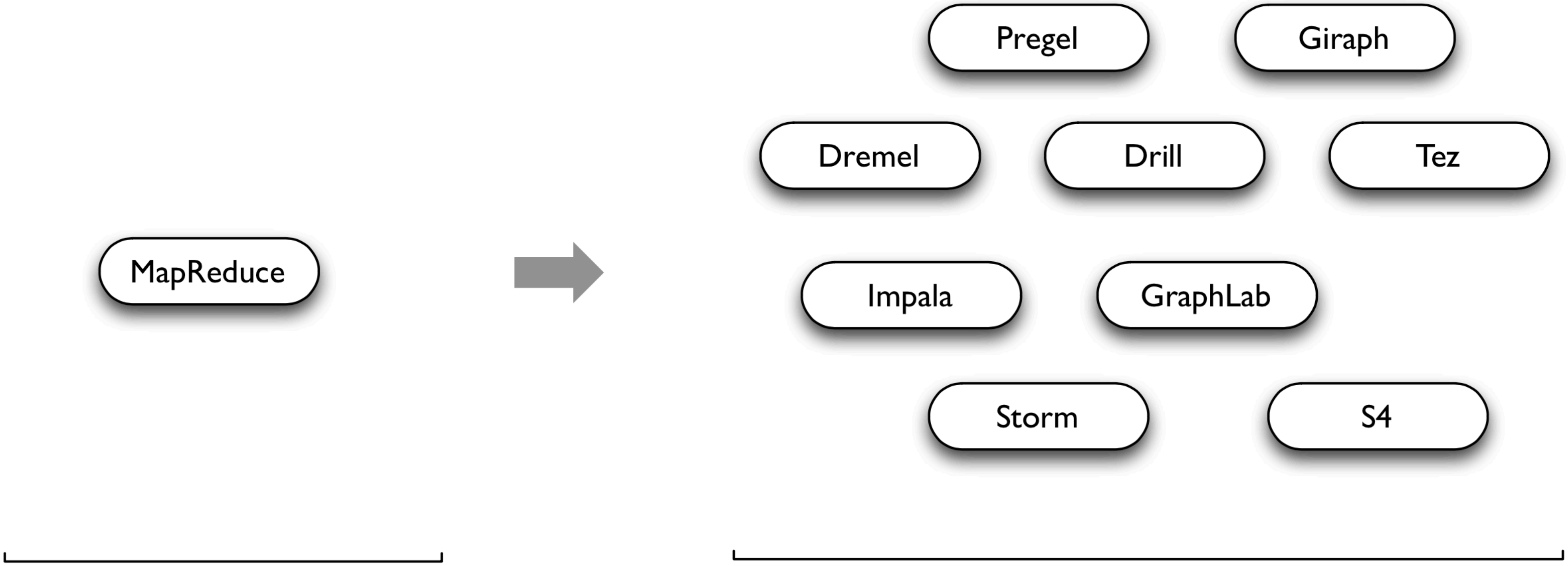
Python:

```python
from operator import add
f = sc.textFile("README.md")
wc = f.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(add)
wc.saveAsTextFile("wc_out.txt")
```

Limitations of MapReduce

- Performance bottlenecks—not all jobs can be cast as batch processes
  - Graphs?
- Programming in Hadoop is hard
  - Boilerplate boilerplate everywhere

Initial Workaround: Specialization



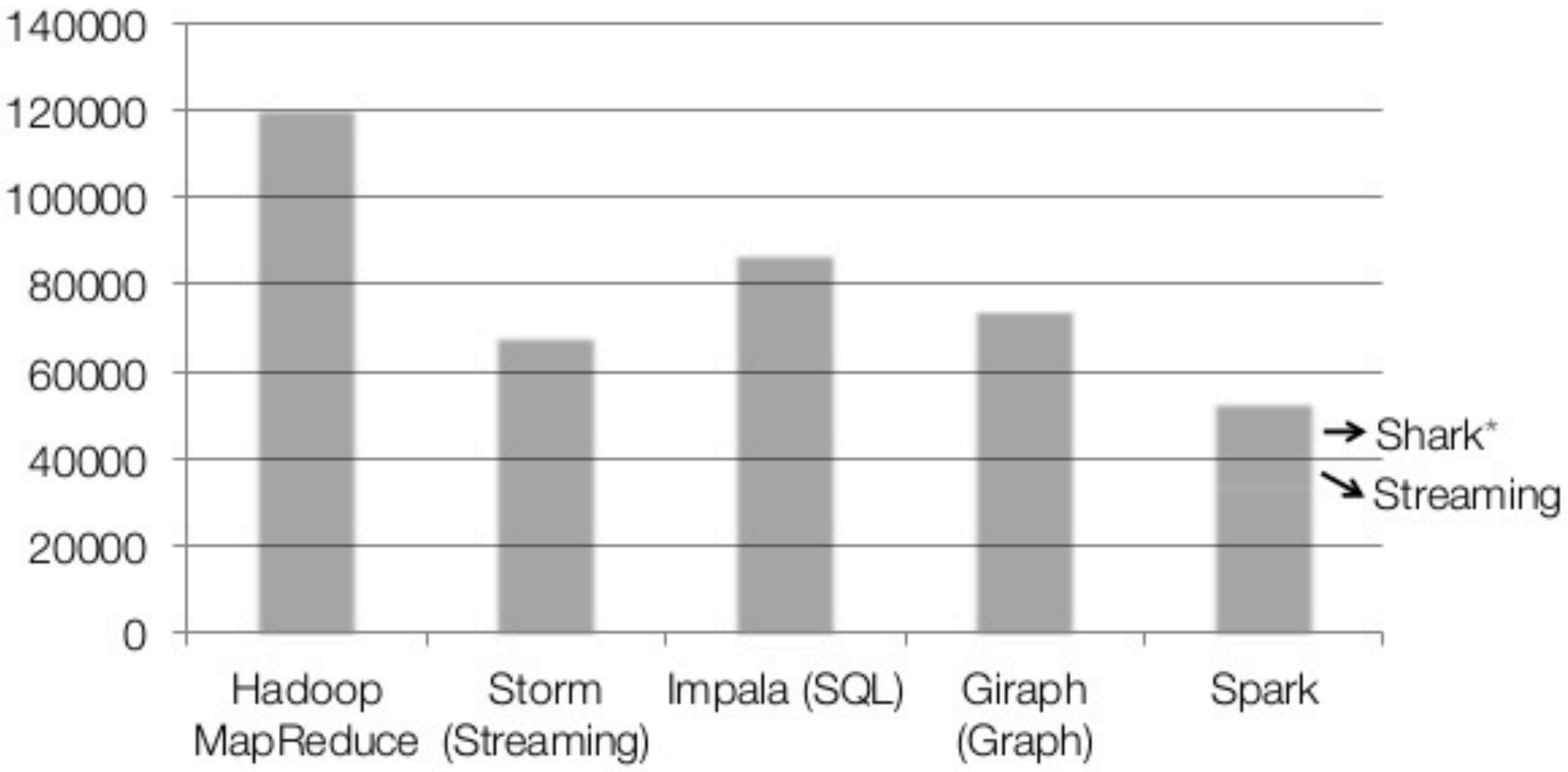
General Batch Processing

Specialized Systems:
iterative, interactive, streaming, graph, etc.

Along Came Spark

- Spark's goal was to *generalize* MapReduce to support new applications within the same engine
- Two additions:
  - Fast data sharing
  - General DAGs (directed acyclic graphs)
- Best of both worlds: easy to program & more efficient engine in general

Codebase Size



non-test, non-example source lines

* also calls into Hive

More on Spark

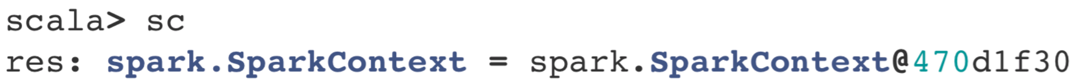
- More general
  - Supports map/reduce paradigm
  - Supports vertex-based paradigm
  - Supports streaming algorithms
  - General compute engine (DAG)
- More API hooks
  - Scala, Java, Python, R
- More interfaces
  - Batch (Hadoop), real-time (Storm), and interactive (???)

Interactive Shells

- Spark creates a SparkSession object (cluster information)
- For either shell: spark
- External programs use a static constructor to instantiate the context
- Pull the SparkContext out via spark.SparkContext

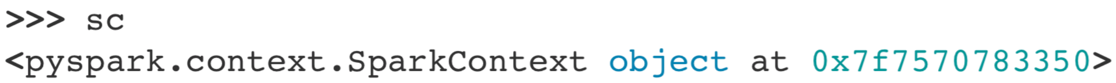
```
./bin/spark-shell

./bin/pyspark
```

Scala:



```
scala> sc
res: spark.SparkContext = spark.SparkContext@470d1f30
```

Python:



```
>>> sc
<pyspark.context.SparkContext object at 0x7f7570783350>
```

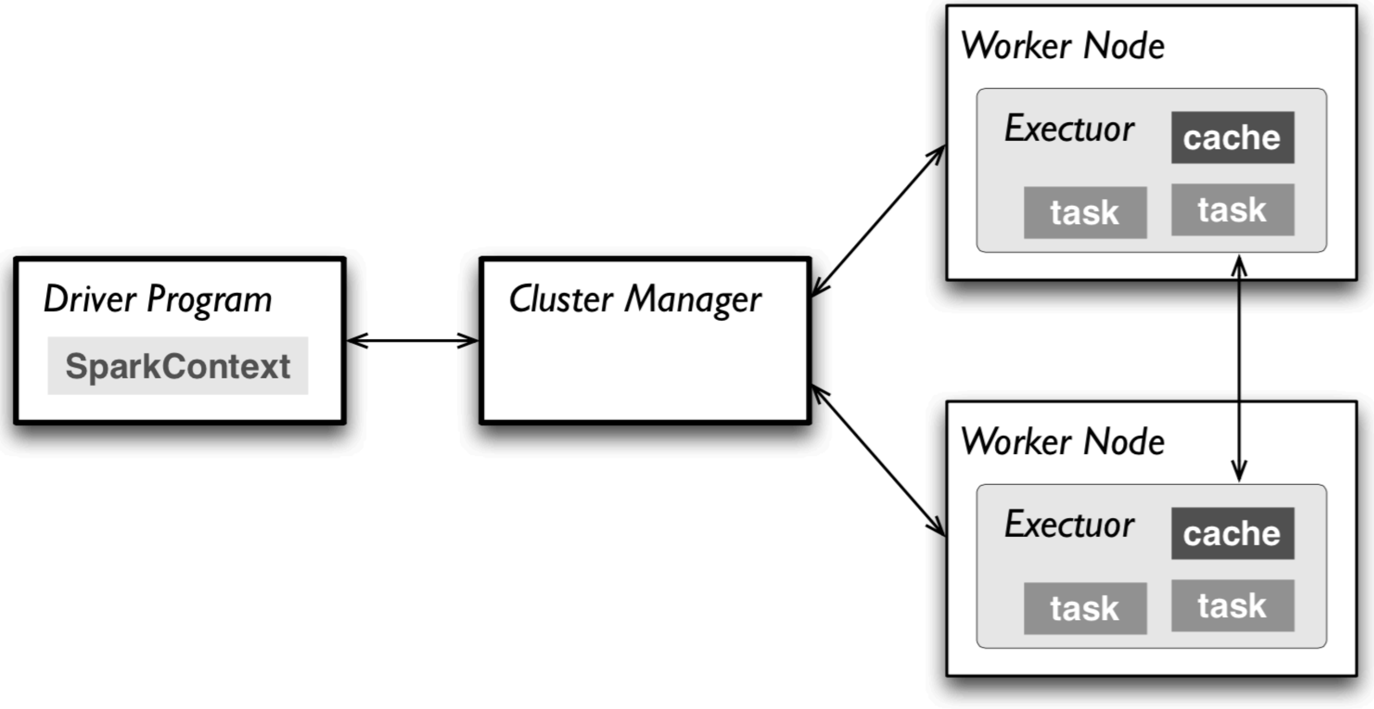
Interactive Shells

- spark-shell --*master*

master	*description*
`local`	run Spark locally with one worker thread (no parallelism)
`local[K]`	run Spark locally with K worker threads (ideally set to # cores)
`spark://HOST:PORT`	connect to a Spark standalone cluster; PORT depends on config (7077 by default)
`mesos://HOST:PORT`	connect to a Mesos cluster; PORT depends on config (5050 by default)

Interactive Shells

- Master connects to the cluster manager, which allocates resources across applications
- Acquires executors on cluster nodes: worker processes to run computations and store data
- Sends app code to executors
- Sends tasks for executors to run

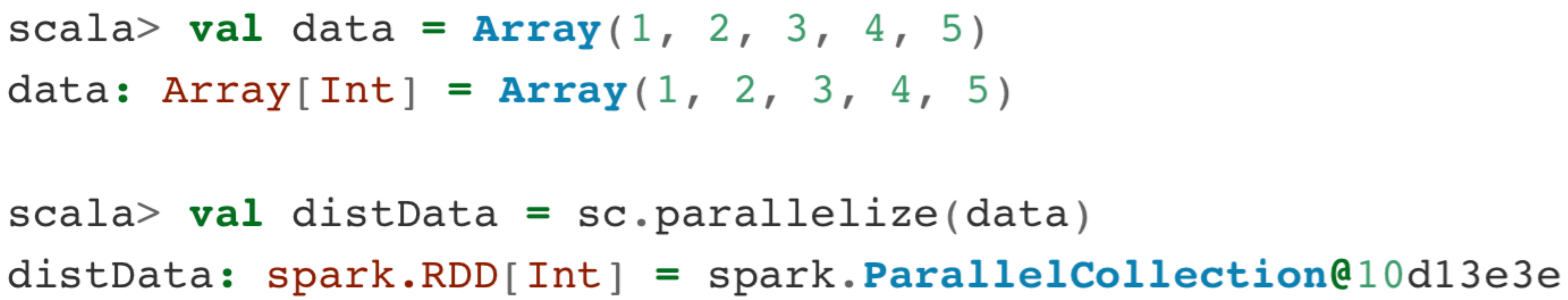


Resilient Distributed Datasets (RDDs)

- **R**esilient **D**istributed **D**atasets (RDDs) are primary data abstraction in Spark
  - Fault-tolerant
  - Can be operated on in parallel
  1. Parallelized Collections
  2. Hadoop datasets
- Two types of RDD operations
  1. Transformations (lazy)
  2. Actions (immediate)

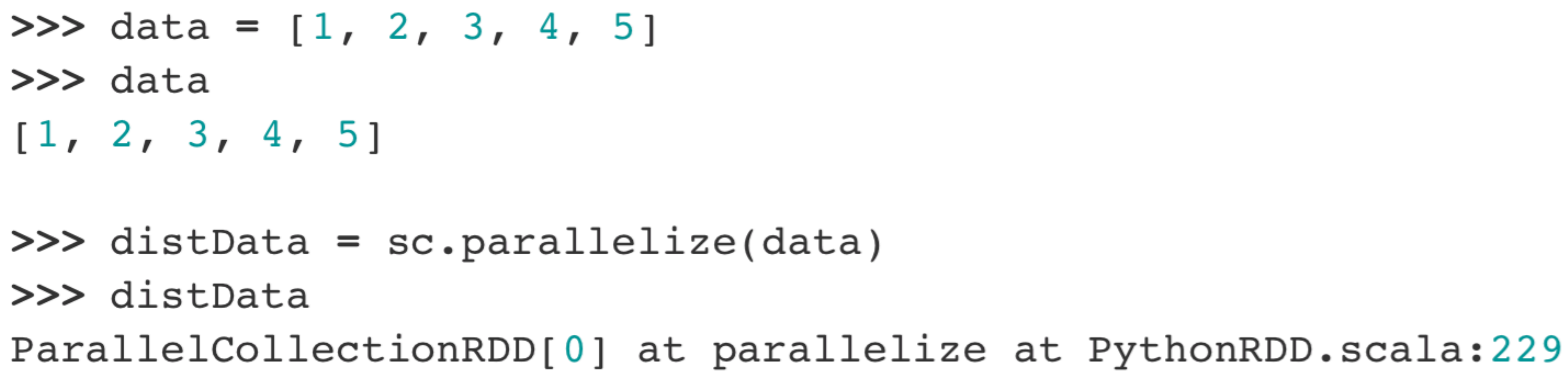
Resilient Distributed Datasets (RDDs)

Scala:



```
scala> val data = Array(1, 2, 3, 4, 5)
data: Array[Int] = Array(1, 2, 3, 4, 5)

scala> val distData = sc.parallelize(data)
distData: spark.RDD[Int] = spark.ParallelCollection@10d13e3e
```

Python:



```
>>> data = [1, 2, 3, 4, 5]
>>> data
[1, 2, 3, 4, 5]

>>> distData = sc.parallelize(data)
>>> distData
ParallelCollectionRDD[0] at parallelize at PythonRDD.scala:229
```

Resilient Distributed Datasets (RDDs)

- Can create RDDs from any file stored in HDFS
  - Local filesystem
  - Amazon S3
  - HBase
- Text files, SequenceFiles, or any other Hadoop `InputFormat`
- Any directory or glob
  - /data/201414*

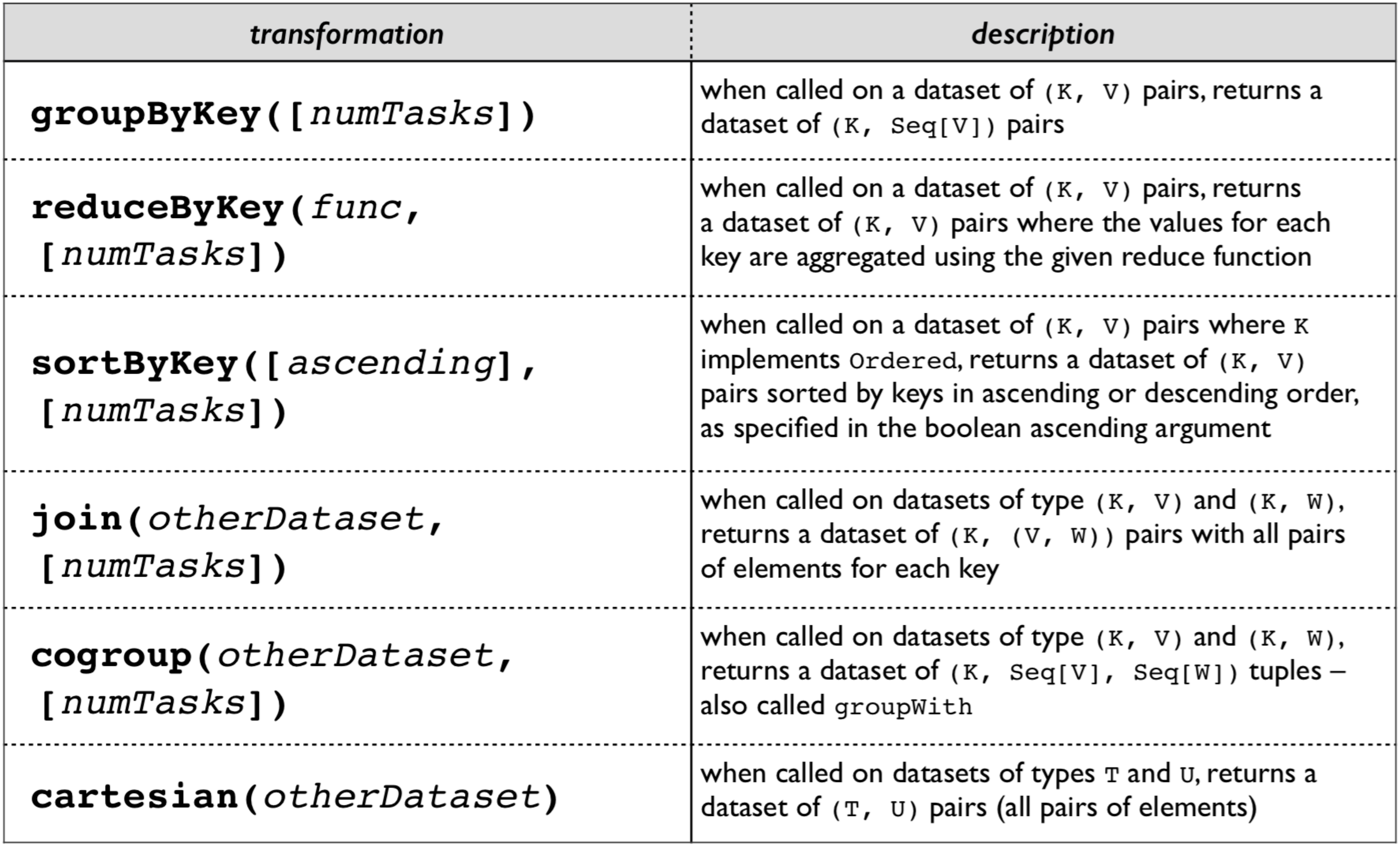
Resilient Distributed Datasets (RDDs)

- Transformations
  - Create a new RDD from an existing one
  - *Lazily* evaluated: results are not immediately computed
    - Pipeline of subsequent transformations can be optimized
    - Lost data partitions can be recovered

Resilient Distributed Datasets (RDDs)

transformation	*description*
map(*func***)**	return a new distributed dataset formed by passing each element of the source through a function *func*
filter(*func***)**	return a new dataset formed by selecting those elements of the source on which *func* returns true
flatMap(*func***)**	similar to map, but each input item can be mapped to 0 or more output items (so *func* should return a Seq rather than a single item)
sample(*withReplacement***,** *fraction***,** *seed***)**	sample a fraction *fraction* of the data, with or without replacement, using a given random number generator *seed*
union(*otherDataset***)**	return a new dataset that contains the union of the elements in the source dataset and the argument
distinct([*numTasks***]))**	return a new dataset that contains the distinct elements of the source dataset

Resilient Distributed Datasets (RDDs)



transformation	*description*
groupByKey([*numTasks*])	when called on a dataset of `(K, V)` pairs, returns a dataset of `(K, Seq[V])` pairs
reduceByKey(*func*, [*numTasks*])	when called on a dataset of `(K, V)` pairs, returns a dataset of `(K, V)` pairs where the values for each key are aggregated using the given reduce function
sortByKey([*ascending*], [*numTasks*])	when called on a dataset of `(K, V)` pairs where `K` implements `Ordered`, returns a dataset of `(K, V)` pairs sorted by keys in ascending or descending order, as specified in the boolean ascending argument
join(*otherDataset*, [*numTasks*])	when called on datasets of type `(K, V)` and `(K, W)`, returns a dataset of `(K, (V, W))` pairs with all pairs of elements for each key
cogroup(*otherDataset*, [*numTasks*])	when called on datasets of type `(K, V)` and `(K, W)`, returns a dataset of `(K, Seq[V], Seq[W])` tuples – also called `groupWith`
cartesian(*otherDataset*)	when called on datasets of types `T` and `U`, returns a dataset of `(T, U)` pairs (all pairs of elements)

Resilient Distributed Datasets (RDDs)

Scala:

```
val distFile = sc.textFile("README.md")
distFile.map(l => l.split(" ")).collect()
distFile.flatMap(l => l.split(" ")).collect()
```

distFile is a collection of lines

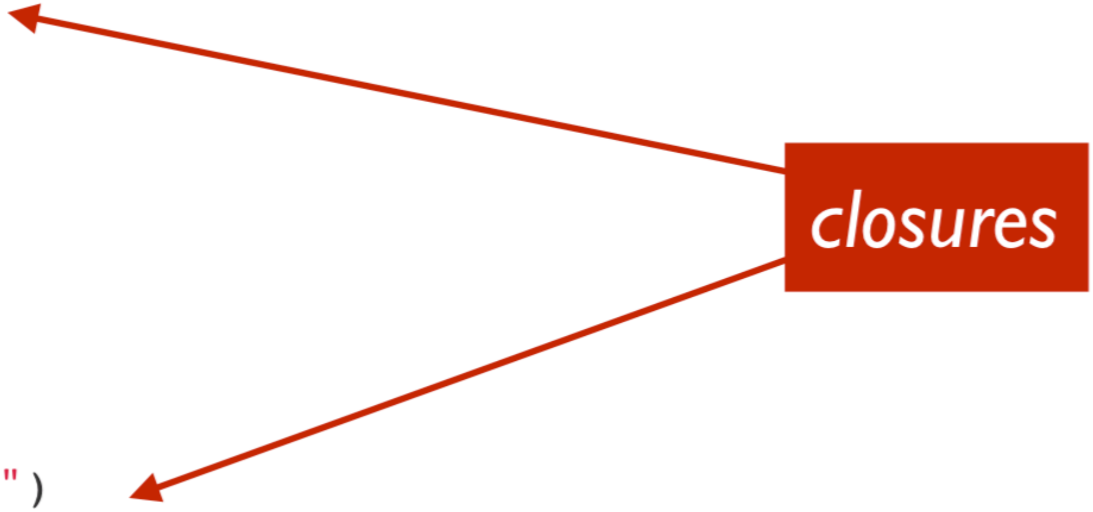
Python:

```
distFile = sc.textFile("README.md")
distFile.map(lambda x: x.split(' ')).collect()
distFile.flatMap(lambda x: x.split(' ')).collect()
```

Resilient Distributed Datasets (RDDs)

Scala:

```
val distFile = sc.textFile("README.md")
distFile.map(l => l.split(" ")).collect()
distFile.flatMap(l => l.split(" ")).collect()
```



closures

Python:

```
distFile = sc.textFile("README.md")
distFile.map(lambda x: x.split(' ')).collect()
distFile.flatMap(lambda x: x.split(' ')).collect()
```

Closures in Java

Java 7:

```
JavaRDD<String> distFile = sc.textFile("README.md");

// Map each line to multiple words
JavaRDD<String> words = distFile.flatMap(
  new FlatMapFunction<String, String>() {
    public Iterable<String> call(String line) {
      return Arrays.asList(line.split(" "));
    }
});
```

Java 8:

```
JavaRDD<String> distFile = sc.textFile("README.md");
JavaRDD<String> words =
    distFile.flatMap(line -> Arrays.asList(line.split(" ")));
```

Resilient Distributed Datasets (RDDs)

- Actions
  - Create a new RDD from an existing one
  - *Eagerly* evaluated: results are immediately computed
    - Applies previous transformations
    - (cache results?)

Resilient Distributed Datasets (RDDs)

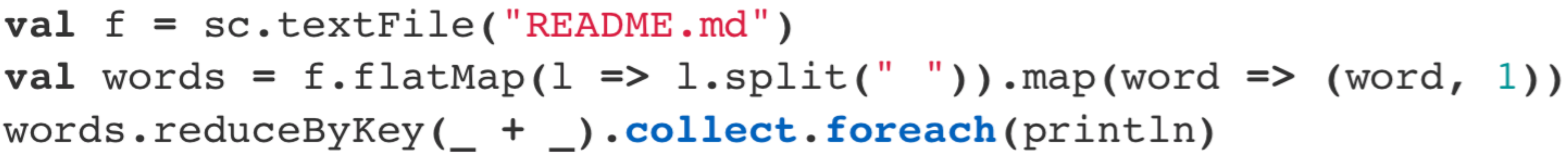
action	*description*
reduce(*func*)	aggregate the elements of the dataset using a function *func* (which takes two arguments and returns one), and should also be commutative and associative so that it can be computed correctly in parallel
collect()	return all the elements of the dataset as an array at the driver program – usually useful after a filter or other operation that returns a sufficiently small subset of the data
count()	return the number of elements in the dataset
first()	return the first element of the dataset – similar to *take(1)*
take(*n*)	return an array with the first *n* elements of the dataset – currently not executed in parallel, instead the driver program computes all the elements
takeSample(*withReplacement*, *fraction*, *seed*)	return an array with a random sample of *num* elements of the dataset, with or without replacement, using the given random number generator seed

Resilient Distributed Datasets (RDDs)

action	*description*
saveAsTextFile(*path*)	write the elements of the dataset as a text file (or set of text files) in a given directory in the local filesystem, HDFS or any other Hadoop-supported file system. Spark will call toString on each element to convert it to a line of text in the file
saveAsSequenceFile(*path*)	write the elements of the dataset as a Hadoop SequenceFile in a given path in the local filesystem, HDFS or any other Hadoop-supported file system. Only available on RDDs of key-value pairs that either implement Hadoop's Writable interface or are implicitly convertible to Writable (Spark includes conversions for basic types like Int, Double, String, etc).
countByKey()	only available on RDDs of type (K, V). Returns a `Map` of (K, Int) pairs with the count of each key
foreach(*func*)	run a function *func* on each element of the dataset – usually done for side effects such as updating an accumulator variable or interacting with external storage systems

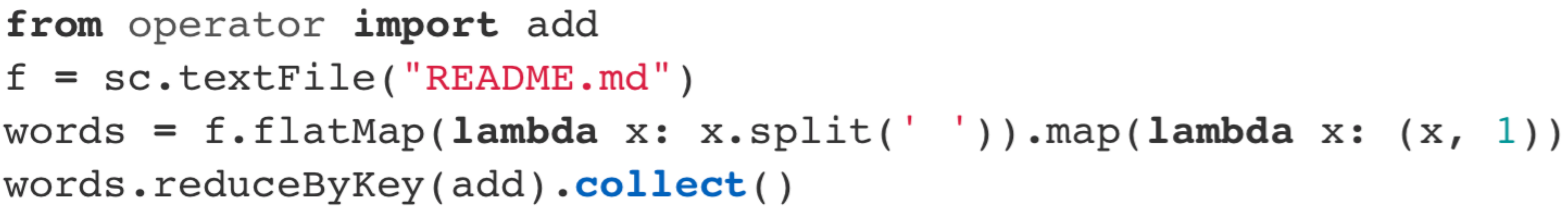
Resilient Distributed Datasets (RDDs)

Scala:



```scala
val f = sc.textFile("README.md")
val words = f.flatMap(l => l.split(" ")).map(word => (word, 1))
words.reduceByKey(_ + _).collect.foreach(println)
```

Python:



```python
from operator import add
f = sc.textFile("README.md")
words = f.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1))
words.reduceByKey(add).collect()
```

Resilient Distributed Datasets (RDDs)

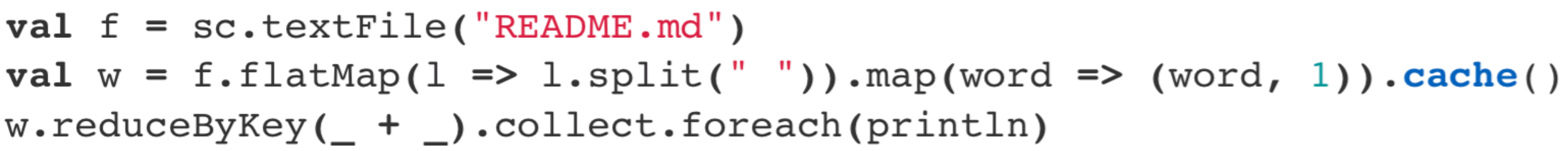
- Spark can persist / cache an RDD in memory across operations
- Each slice is persisted in memory and reused in subsequent actions involving that RDD
- Cache provides fault-tolerance: if partition is lost, it will be recomputed using transformations that created it

Resilient Distributed Datasets (RDDs)

transformation	*description*
MEMORY_ONLY	Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, some partitions will not be cached and will be recomputed on the fly each time they're needed. This is the default level.
MEMORY_AND_DISK	Store RDD as deserialized Java objects in the JVM. If the RDD does not fit in memory, store the partitions that don't fit on disk, and read them from there when they're needed.
MEMORY_ONLY_SER	Store RDD as serialized Java objects (one byte array per partition). This is generally more space-efficient than deserialized objects, especially when using a fast serializer, but more CPU-intensive to read.
MEMORY_AND_DISK_SER	Similar to MEMORY_ONLY_SER, but spill partitions that don't fit in memory to disk instead of recomputing them on the fly each time they're needed.
DISK_ONLY	Store the RDD partitions only on disk.
MEMORY_ONLY_2, **MEMORY_AND_DISK_2,** etc	Same as the levels above, but replicate each partition on two cluster nodes.

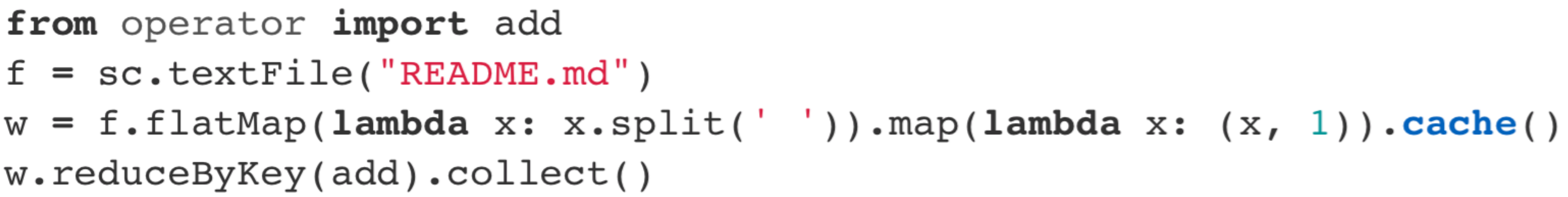
Resilient Distributed Datasets (RDDs)

Scala:



```scala
val f = sc.textFile("README.md")
val w = f.flatMap(l => l.split(" ")).map(word => (word, 1)).cache()
w.reduceByKey(_ + _).collect.foreach(println)
```

Python:



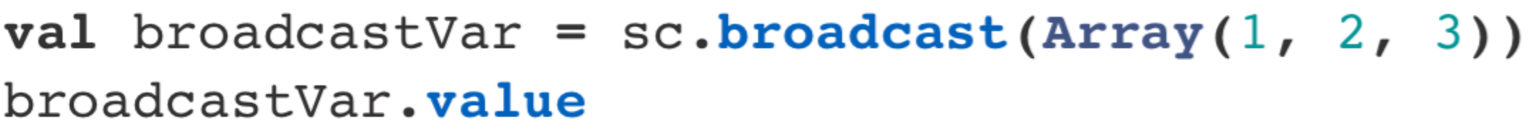
```python
from operator import add
f = sc.textFile("README.md")
w = f.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).cache()
w.reduceByKey(add).collect()
```

Broadcast Variables

- Spark's version of Hadoop's DistributedCache
- Read-only variable cached on each node
- Spark [internally] distributed broadcast variables in such a way to minimize communication cost

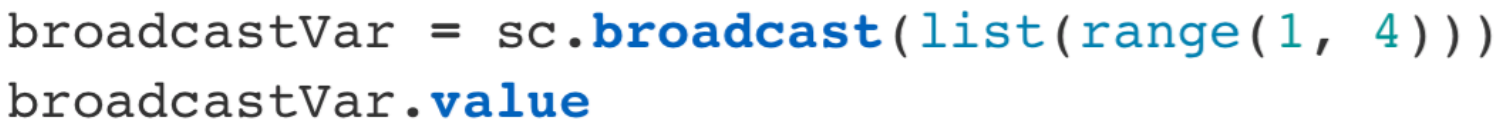
Broadcast Variables

Scala:



```scala
val broadcastVar = sc.broadcast(Array(1, 2, 3))
broadcastVar.value
```

Python:



```python
broadcastVar = sc.broadcast(list(range(1, 4)))
broadcastVar.value
```

Accumulators

- Spark's version of Hadoop's Counter
- Variables that can only be added through an associative operation
- Native support of numeric accumulator types and standard mutable collections
  - Users can extend to new types
- Only driver program can *read* accumulator value

Accumulators

Scala:

```
val accum = sc.accumulator(0)
sc.parallelize(Array(1, 2, 3, 4)).foreach(x => accum += x)

accum.value
```

driver-side

Python:

```
accum = sc.accumulator(0)
rdd = sc.parallelize([1, 2, 3, 4])
def f(x):
   global accum
   accum += x

rdd.foreach(f)

accum.value
```

Key/Value Pairs

Scala:

```
val pair = (a, b)

    pair._1 // => a
    pair._2 // => b
```

Python:

```
pair = (a, b)

    pair[0] # => a
    pair[1] # => b
```

Java:

```
Tuple2 pair = new Tuple2(a, b);

    pair._1 // => a
    pair._2 // => b
```

Resources

- Original slide deck: http://cdn.liber118.com/workshop/itas_workshop.pdf
- Code samples:
  - https://gist.github.com/ceteri/f2c3486062c9610eac1d
  - https://gist.github.com/ceteri/8ae5b9509a08c08a1132
  - https://gist.github.com/ceteri/11381941

Questions?