Final Project M2

CS 327E
November 6, 2017

1) Consider a key-value pair RDD. Which RDD operation is used to switch the keys and values in this type of RDD?

A) map()
B) collect()
C) union()
D) mapValues()

2) Consider a key-value pair RDD. Which RDD operation is used to sort the keys in this type of RDD?

A) reduceByKey()
B) sortByKey()
C) sortByValue()
D) groupByKey()

3) What is the value of `y`?

```
x = sc.parallelize([2, 4, 1, 7])
y = x.max()
```

A) [2, 4, 1, 7]
B) [1, 2, 4, 7]
C) 7
D) 14

4) Consider the action `count()` which counts the number of elements of an RDD. What is the output of `count()` in the code sample below?

```
>>> text = sc.textFile("file.txt")
>>> text.collect()
[u'Sam I am.', u'Sam I am.', u'That Sam I am.', u'I do not like
 that Sam I am!']
>>> words = text.map(lambda x: x.split(" "))
>>> words.collect()
[[u'Sam', u'I', u'am.'], [u'Sam', u'I', u'am.'], [u'That', u'Sa
m', u'I', u'am.'], [u'I', u'do', u'not', u'like', u'that', u'Sa
m', u'I', u'am!']]
>>> words.count()
```

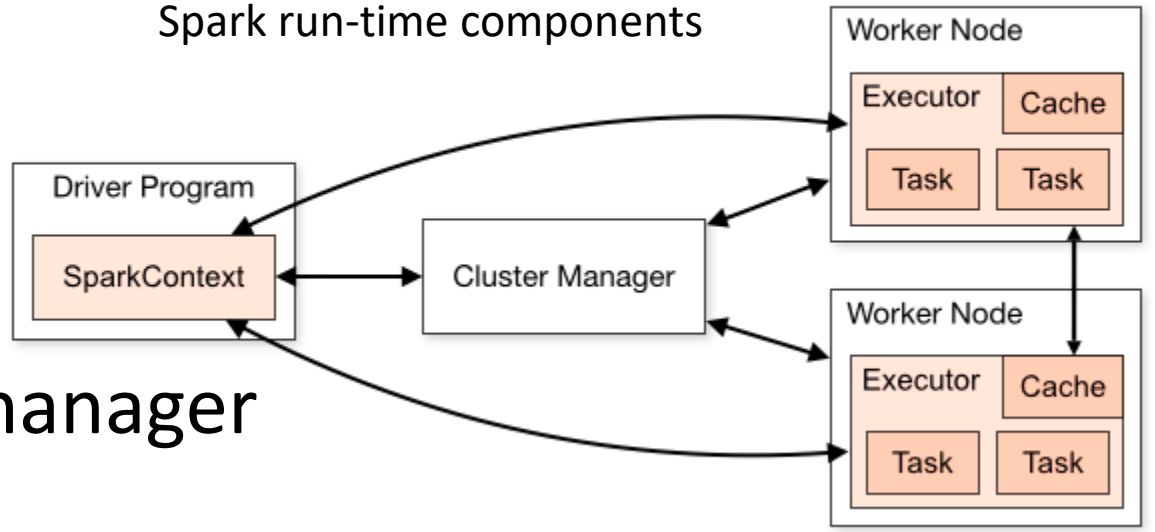
A) 4

B) 18

5) `Broadcast` variables and `accumulators` are two types of global variables in Spark. `Broadcast` variables are used when executors need to read some shared data whereas `accumulators` are used when the executors need to update some shared data.

A) True
B) False

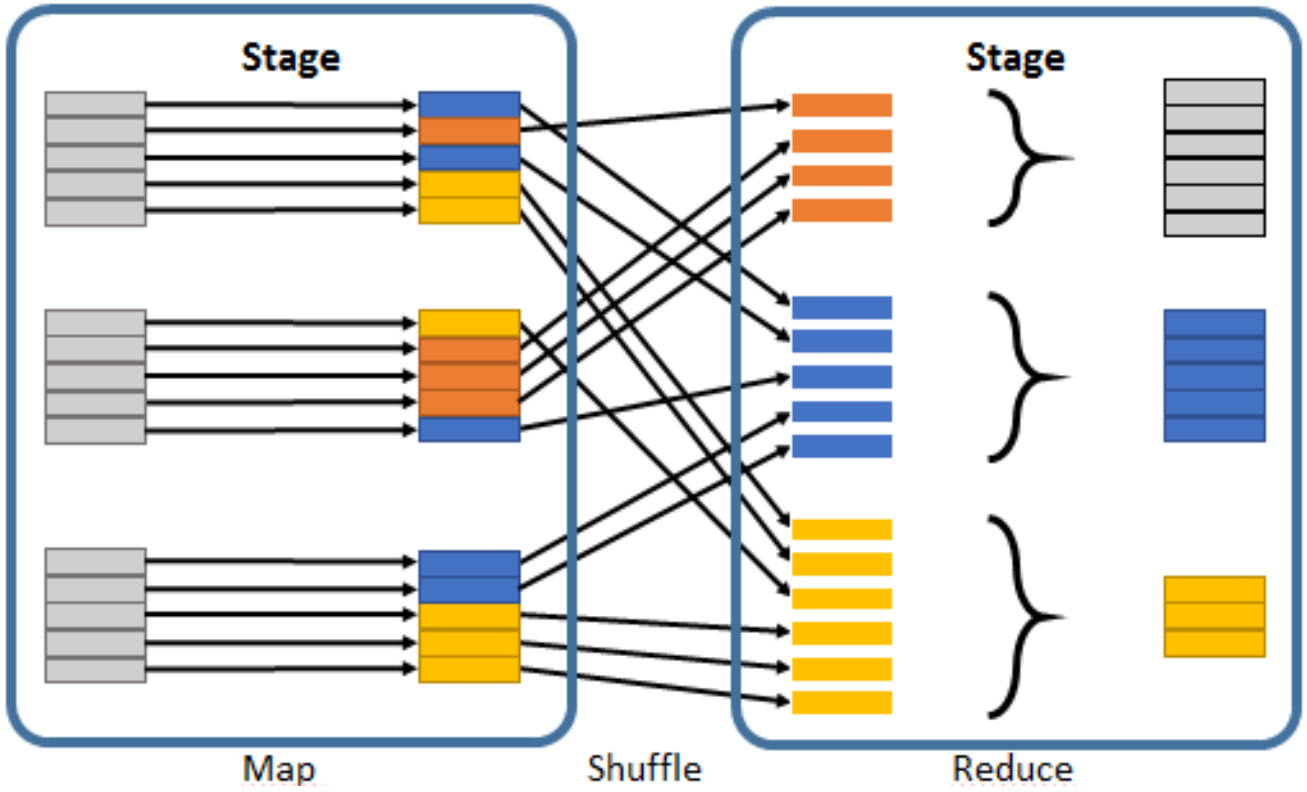
Spark Architecture



- Key components:
  Driver, executors, and cluster manager
- Driver sends tasks to executors and collects the results
- Executors accept tasks from driver, execute the tasks, and return the results back to the driver
- Cluster manager schedules the tasks and manages their resources

Job Stages

- RDDs are partitioned across executors
- Re-partitioning = sending data to a new executor across the cluster (aka shuffle)
- Shuffle phase is slow due to network traffic
- Shuffles determine the number of stages of a job



Job Submission Options

Deploys job on cluster:

```
$ spark-submit <file>.py
```

Deploys job locally:

```
$ spark-submit --master local <file>.py
```

Deploys job on "local" cluster:

```
$ spark-submit --master local[*] <file>.py
```

Milestone 2

http://www.cs.utexas.edu/~scohen/projects/m2-assignment.pdf