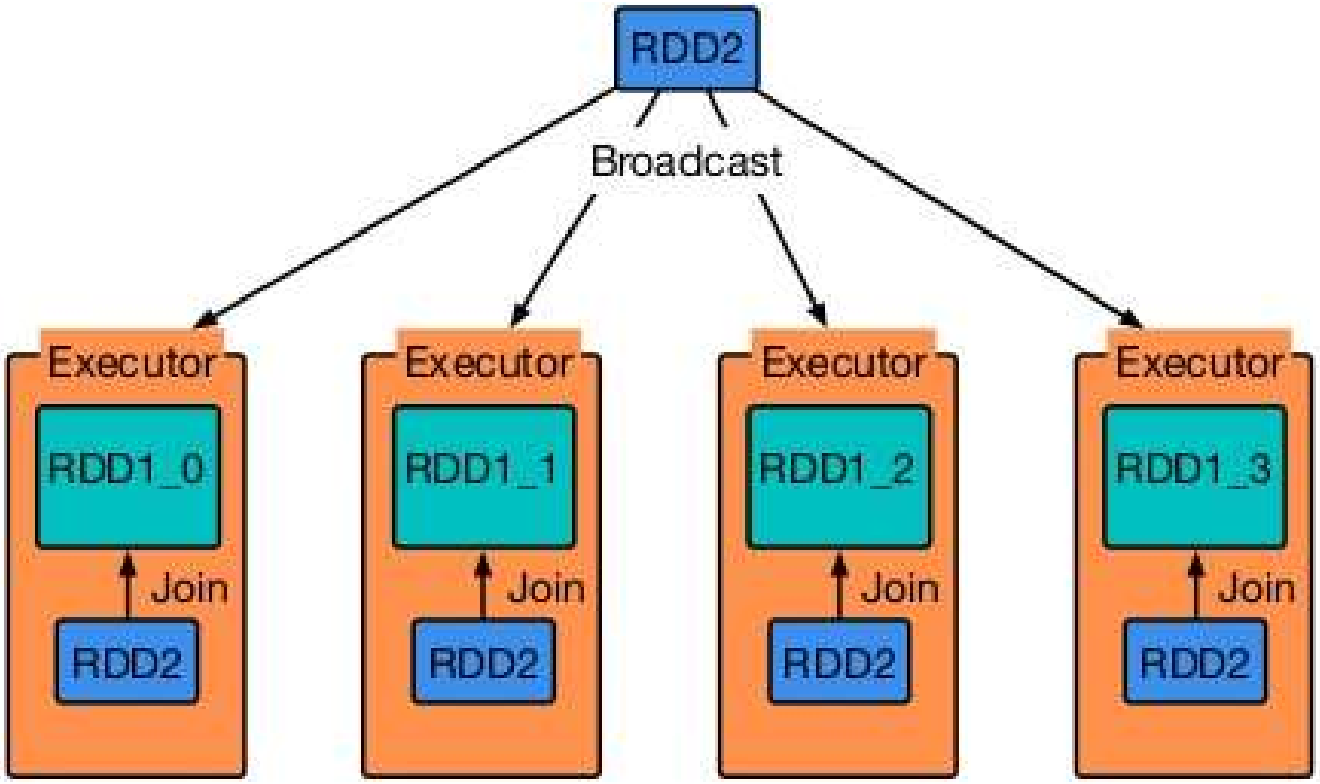
Apache Spark

Using Broadcast Variables and Accumulators

Broadcast Variables

- Is read-only

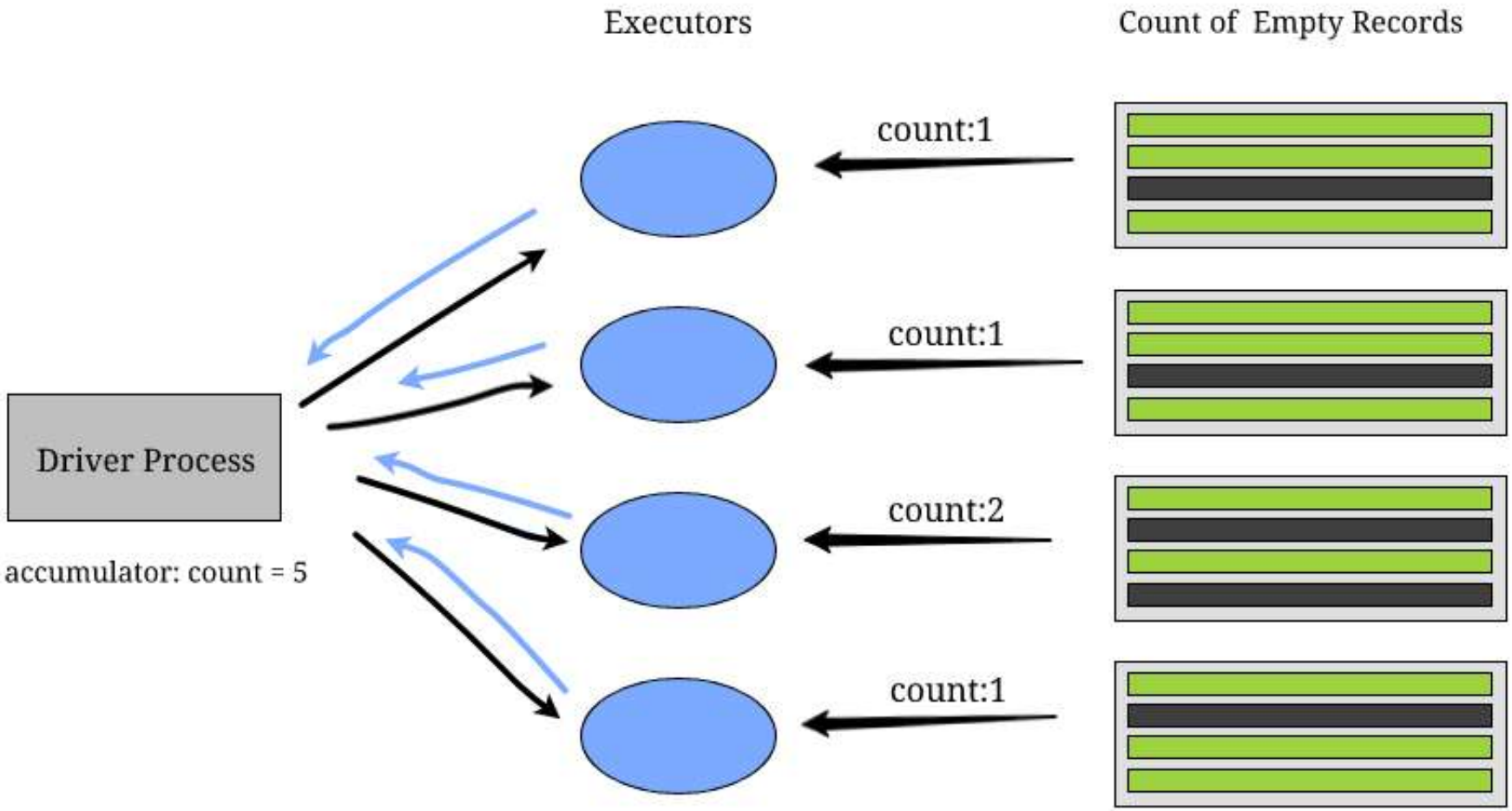


Broadcast Variables

```
val data = spark.textFile(...).map(readPoint).cache()

// Random Projection
Val M = spark.broadcast(Matrix.random(N))

var w = Vector.random(D)

for (i <- 1 to ITERATIONS) {
  val gradient = data.map(p =>
    (1 / (1 + exp(-p.y*(w.dot(p.x.dot(M.value))))) - 1)
      * p.y * p.x
  ).reduce(_ + _)
  w -= gradient
}

println("Final w: " + w)
```

Solution: mark M as broadcast variable

Accumulator



Accumulator

```scala
import org.apache.spark.{SparkConf, SparkContext}

object Add1 {
  def main(args: Array[String]): Unit = {
    val conf = new SparkConf().setMaster("local[*]").setAppName("add1")
    val sc = new SparkContext(conf)
    val list1 = List(30, 50, 70, 60, 10, 20)
    val rdd1 = sc.parallelize(list1, numSlices = 2)

    val acc = sc.longAccumulator
    val rdd2 = rdd1.map(x => {
      acc.add(1)
      x
    })
    rdd2.collect()
    println(acc.value)
    sc.stop()

  }
}
```

Thanks