DLI Accelerated Data Science Teaching Kit

# Lecture 11.2 - Example Spark Programs

# Example: Log Mining

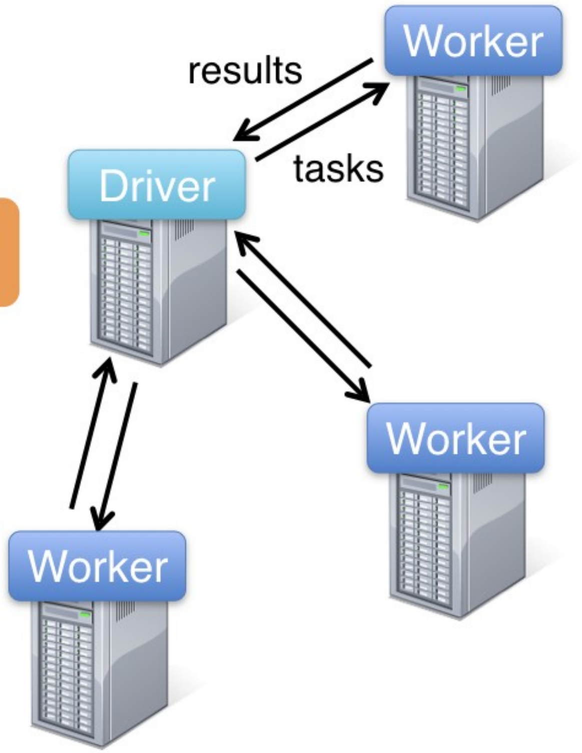
Load error messages from a log into memory, then interactively search for various patterns

```
lines = spark.textFile(“hdfs://...”)
errors = lines.filter(_.startsWith(“ERROR”))
messages = errors.map(_.split(‘\t’)(2))
cachedMsgs = messages.cache()

cachedMsgs.filter(_.contains(“foo”)).count
cachedMsgs.filter(_.contains(“bar”)).count
. . .
```

Base RDD

Transformed RDD

Action



**Result:** scaled to 1 TB data in 5-7 sec (vs 170 sec for on-disk data)

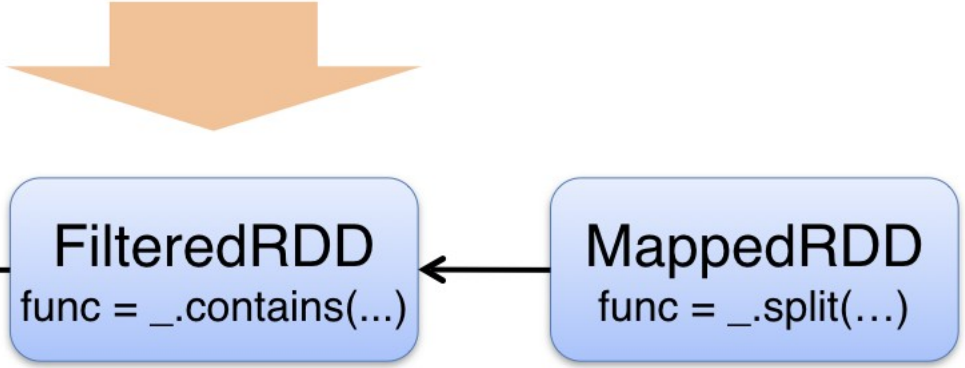
http://ananthakumaran.in/2010/03/29/scala-underscore-magic.html
http://www.slideshare.net/normation/scala-dreaded

# Fault Tolerance

RDDs track the series of transformations used to build them (their *lineage*) to recompute lost data

E.g:

```
messages = textFile(...).filter(_.contains(“error”))
                        .map(_.split(‘\t’)(2))
```

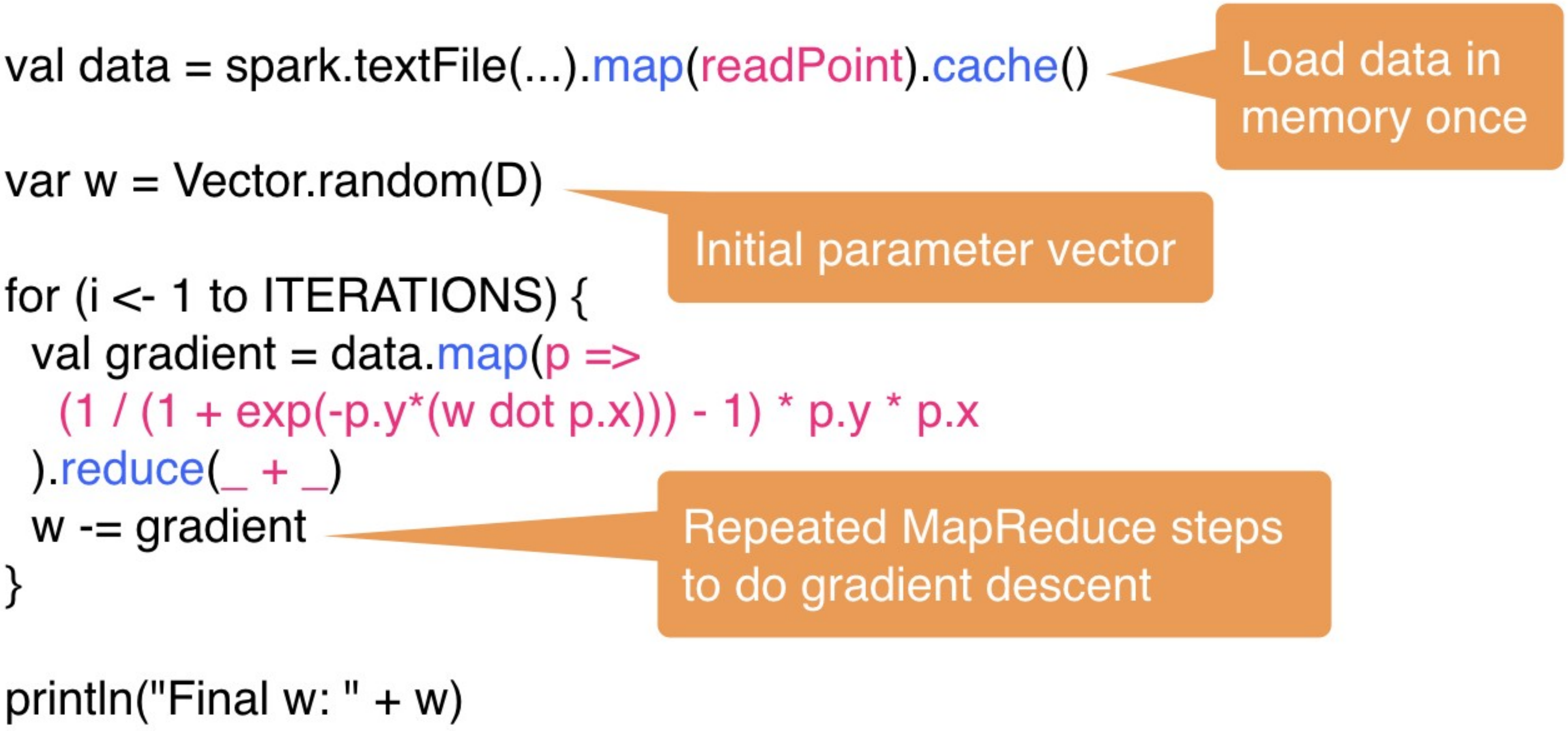


HadoopRDD
path = hdfs://…

FilteredRDD
func = _.contains(...)

MappedRDD
func = _.split(…)

# Example: Logistic Regression



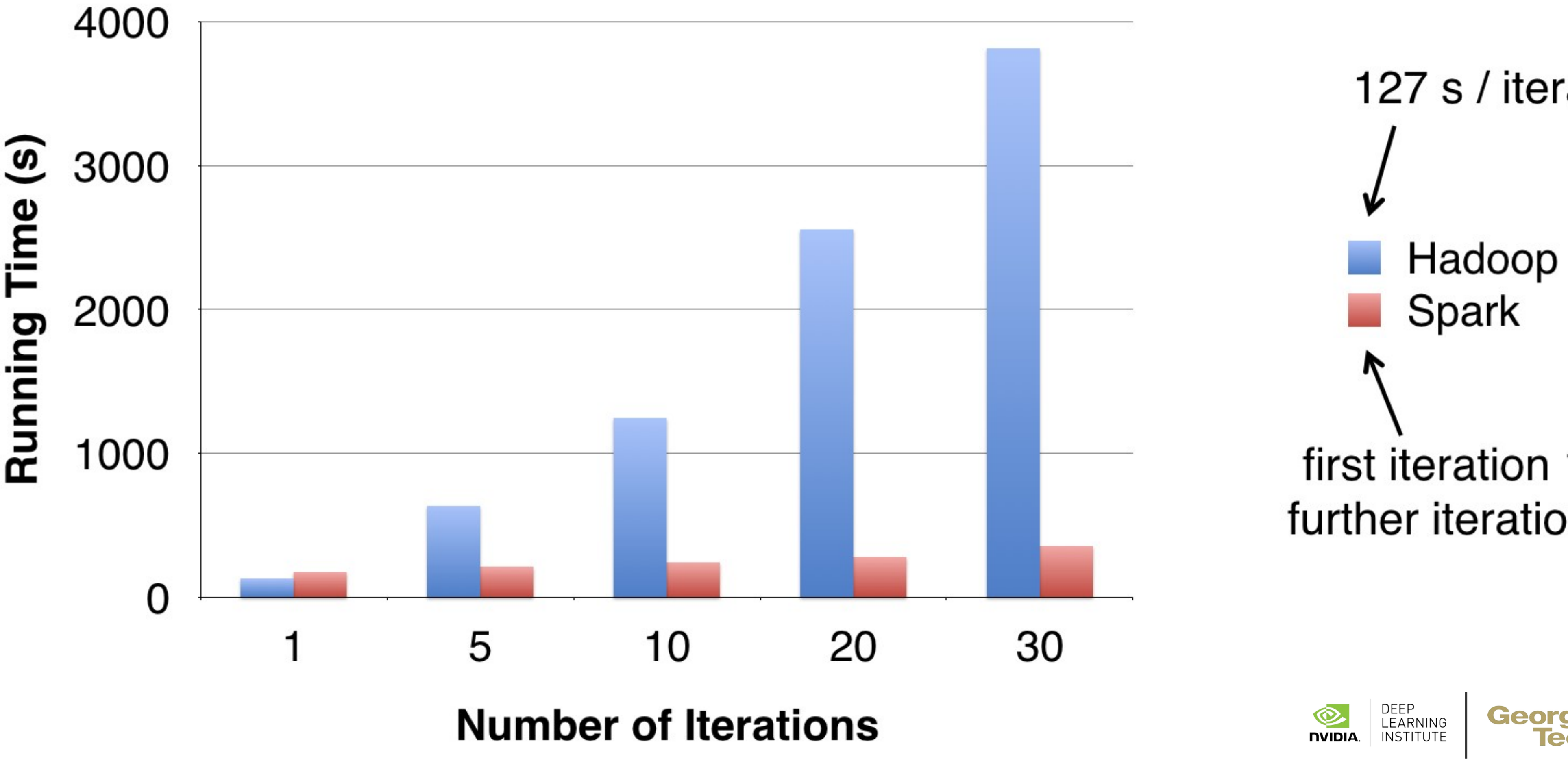
```
val data = spark.textFile(...).map(readPoint).cache()

var w = Vector.random(D)

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

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

Load data in memory once

Initial parameter vector

Repeated MapReduce steps to do gradient descent

# Logistic Regression Performance



127 s / iteration

first iteration 174 s
further iterations 6 s

DLI Accelerated Data Science Teaching Kit

# Thank You

We thank Dr. Matei Zaharia for sharing teaching materials for Spark.