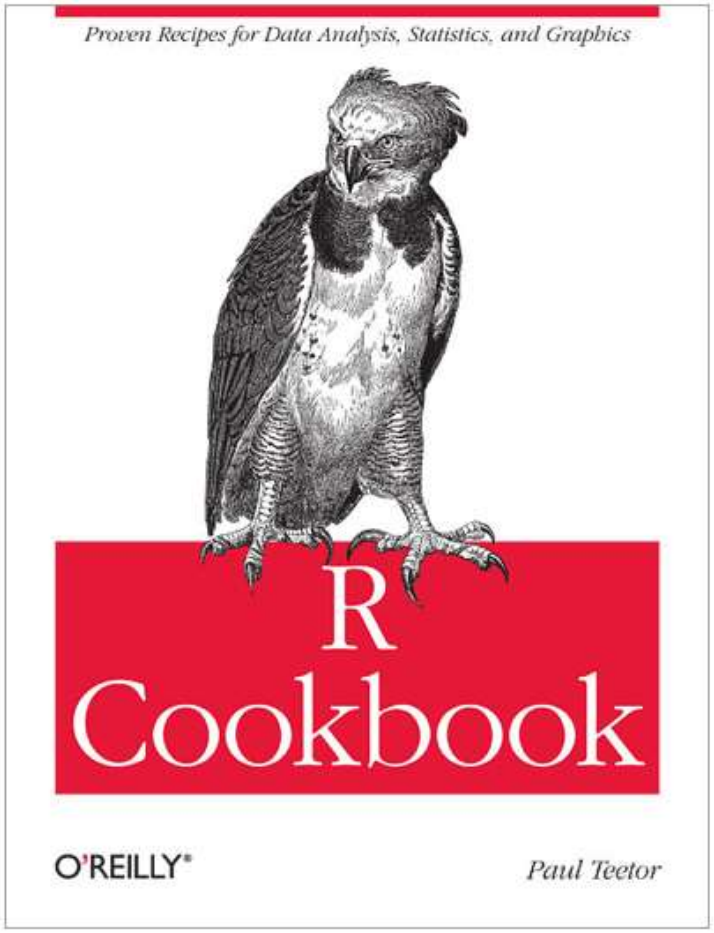
james j.d. long
jdlong@gmail.com

december 6, 2011

london R user group



"With 95% confidence, I fail to reject that R Cookbook *is the best book for learning and using the important stats functions in R."*

—JD Long

stackoverflow

Top r Askers

Last 30 Days

52	24	John	866	1 • 9
51	9	Tyler Rinker	570	1 • 18
31	8	SFun28	2,756	4 • 19
28	6	songpants	269	8
27	4	JD Long	11k	3 • 27 • 82

All Time

455	86	JD Long	11k	3 • 27 • 82
306	90	SFun28	2,756	4 • 19
236	81	Tal Galili	1,974	7 • 36
223	88	Brandon Bertelse	4,825	5 • 29
214	34	gsk3	5,736	11 • 38

SEGUE

parallel R
in the cloud
two lines of code

no kidding!

Data Analysis in the Distributed World

Parallel R

O'REILLY®

Q. Ethan McCallum & Stephen Weston

SEGUE

why...

so i've got this problem...

reinsurance simulations

updated frequently for one month

on my laptop...

- each sim takes ~ 1 min
- 10k sims * 1 min = ~ 7 days

no need for full map/reduce

embarrassingly parallel

SEGUE

you've seen ”word count” demos...

segue has nothing to do with that

big cpu, not big data

my options...

SEGUE

- make the code faster
- build a cluster
  - type
    - snow
    - mpi
    - hadoop
  - location
    - self hosted
    - amazon web services
      - ec2
      - emr
    - rackspace

SEGUE

syntax...

```
require(segue)
myCluster <- createCluster()
```

contratulations. we've built a cluster!

SEGUE

more syntax...

parallel apply() on lists:

base R:

```
lapply( X, FUN, … )
```

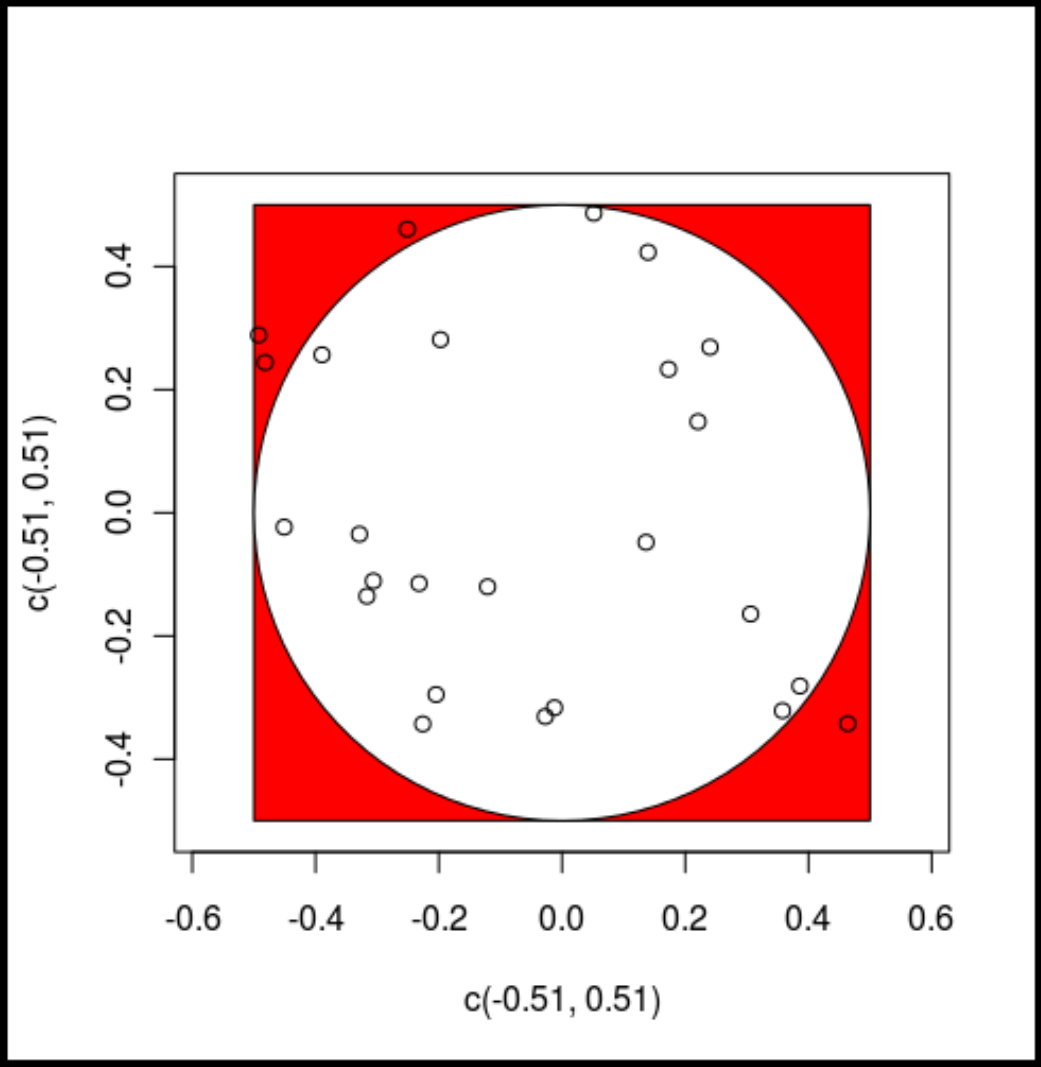
segue:

```
emrlapply( clusterObject, X, FUN, … )
```

SEGUE

example...

stochastic pi simulation (again)



SEGUE

example...

```
estimatePi <- function( seed ){
  set.seed(seed)
  numDraws <- 1000000
  r <- .5
  x <- runif(numDraws, min=-r, max=r)
  y <- runif(numDraws, min=-r, max=r)
  inCircle <- ifelse( (x^2 + y^2)^.5 < r , 1, 0)
  return(sum(inCircle) / length(inCircle) * 4)
}
```

```
seedList <- as.list(1:1000)
        require(segue)
        myCluster <- createCluster(20)
        myEstimates <- emrlapply( myCluster, seedList, estimatePi )
        stopCluster(myCluster)
        myPi <- Reduce(sum, myEstimates) / length(myEstimates)
        format(myPi, digits=10)
```

https://gist.github.com/764370

howzit work?

SEGUE

createCluster()

cluster object:
list of parameters

temp dirs:
local
S3 for EMR

bootstrap:
update R
update packages

~ 10-15 minutes

howzit work?

SEGUE

emrlapply()

list is serialized to CSV and uploaded to S3 – streaming input file

function, arguments, r objects, etc are saved & uploaded

EMR copies files to nodes – mapper.R picks them up

CSV is input to mapper.R applies function to each list element

output is serialized into emr part-xxxxx files on s3

part files are downloaded to R and deserialized

deserialized results are reordered and put into a list object

SEGUE

options?

numInstances	number of ec2 machines to fire up
cranPackages	cran packages to load on each cluster node
filesOnNodes	files to be loaded on each node
rObjectsOnNodes	R objects to put on the worker nodes
enableDebugging	start emr debugging
instancesPerNode	number of R instances per node
masterInstanceType	ec2 instance type for the master node
slaveInstanceType	ec2 instance type for the slave nodes
location	ec2 location name for the cluster
ec2KeyName	ec2 key used for logging into the main node
copy.image	copy the entire local environment to the nodes?
otherBootstrapActions	other bootstrap actions to run
sourcePackagesToInstall	R source packages to be installed on each node

https://gist.github.com/764370

SEGUE

when to use segue...

embarrassingly parallel

cpu bound

apply on lists with many items

object size: to / from s3 roundtrip

each job has a fixed & marginal cost

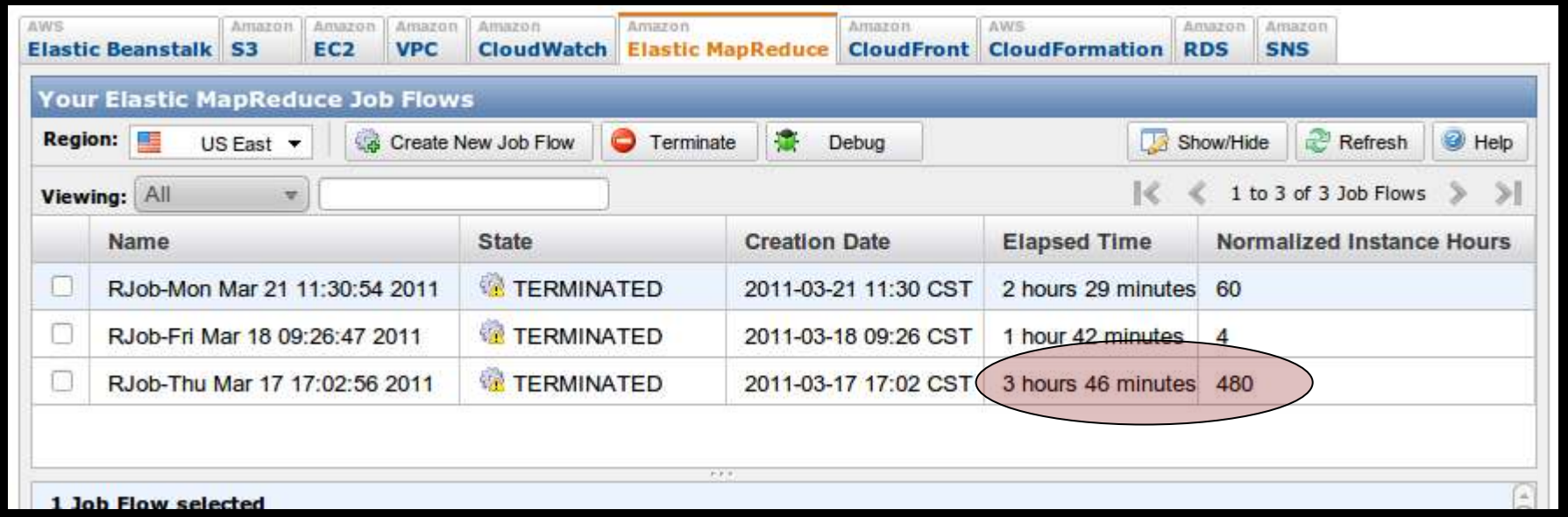
SEGUE

downside of segue...

embarrassingly parallel failure

SEGUE

reasons daddy drinks...
(a.k.a things vendors never say)
keep one eye on aws dashboard



united nations considers debugging of segue jobs ”torture” under geneva convention

more reasons daddy drinks...

SEGUE

if you use segue you will see:

- unreproducable errors
- clusters that never start
- temp buckets in your s3 acct
- clusters left running
- i/o that takes longer than calcs

but... i've never had a ”wrong” answer

SEGUE

Immediate segue future...

maintenance issues:

R releases change

emr changes

vendor lock-in to amazon

whirr as solution?

foreach %dopar% backend?

SEGUE

imagine the future...

R objects backed by clusters
as.hdfs.data.frame(data)

operations converted to map reduce jobs transparently

other abstractions...

SEGUE

segue project page
http://code.google.com/p/segue/

google groups
http://groups.google.com/group/segue-r

see also...

rhipe – program m/r in R
http://www.stat.purdue.edu/~sguha/rhipe/

revolution analytics rhadoop:
https://github.com/RevolutionAnalytics/RHadoop/wiki