SEGUE

parallel R

in the cloud

two lines of code

no kidding!

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

SEGUE

my options...

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

lowest startup costs

SEGUE

what did my mind's eye see?

SEGUE

”simply irresistible” ~8000 hp

SEGUE

syntax...

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

contratulations. we've built a hadoop cluster!

SEGUE

more syntax...

parallel apply() on lists:

base R:

```
lapply( X, FUN, ... )
```

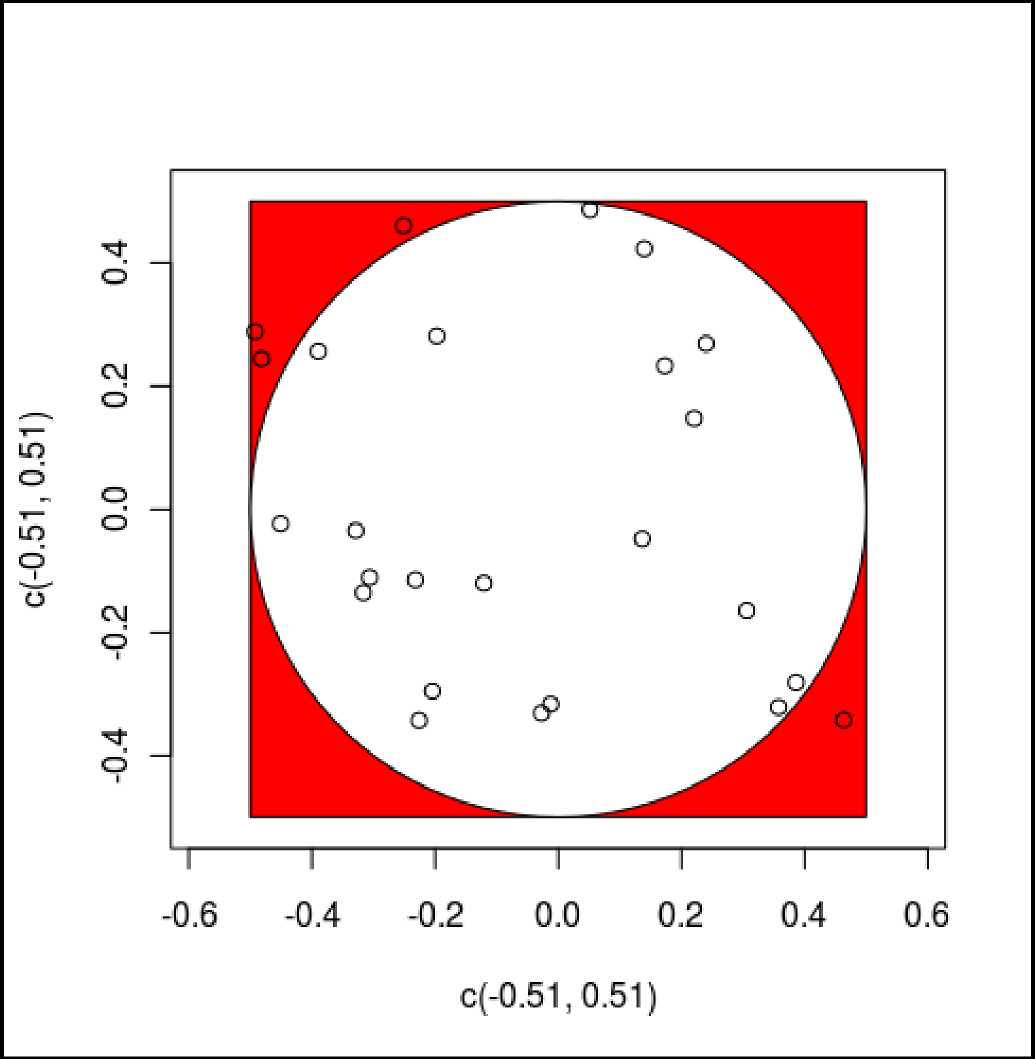
segue:

```
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

```
createCluster( numInstances=2,
               cranPackages,
               filesOnNodes,
               rObjectsOnNodes,
               enableDebugging=FALSE,
               instancesPerNode,
               masterInstanceType="m1.small",
               slaveInstanceType="m1.small",
               location="us-east-1a",
               ec2KeyName,
               copy.image=FALSE,
               otherBootstrapActions,
               sourcePackagesToInstall)
```

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

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

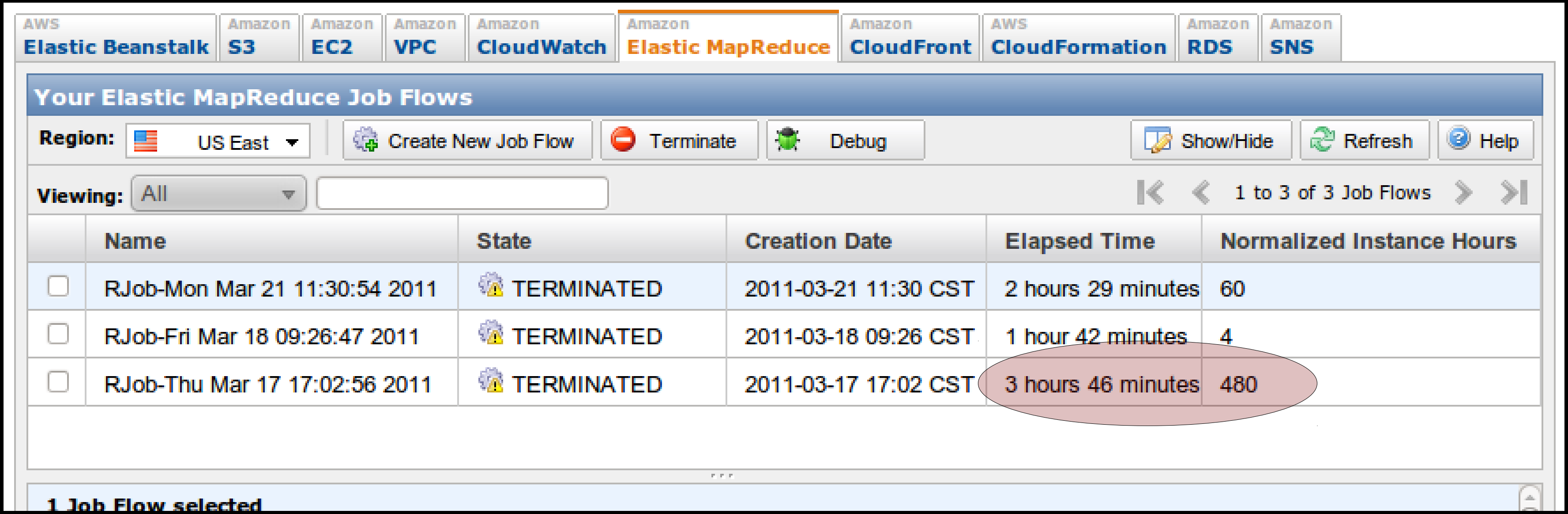
downside of segue...
embarrassingly parallel failure

SEGUE

SEGUE

reasons daddy drinks...
a.k.a things vendors never say

keep one eye on aws dashboard



Your Elastic MapReduce Job Flows

Region: US East

Viewing: All

1 to 3 of 3 Job Flows

	Name	State	Creation Date	Elapsed Time	Normalized Instance Hours
☐	RJob-Mon Mar 21 11:30:54 2011	TERMINATED	2011-03-21 11:30 CST	2 hours 29 minutes	60
☐	RJob-Fri Mar 18 09:26:47 2011	TERMINATED	2011-03-18 09:26 CST	1 hour 42 minutes	4
☐	RJob-Thu Mar 17 17:02:56 2011	TERMINATED	2011-03-17 17:02 CST	3 hours 46 minutes	480

1 Job Flow selected

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

imediate 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

abstractions...

SEGUE

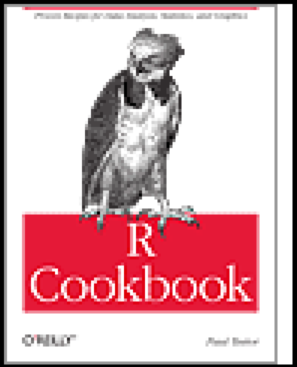
segue project page
http://code.google.com/p/segue/

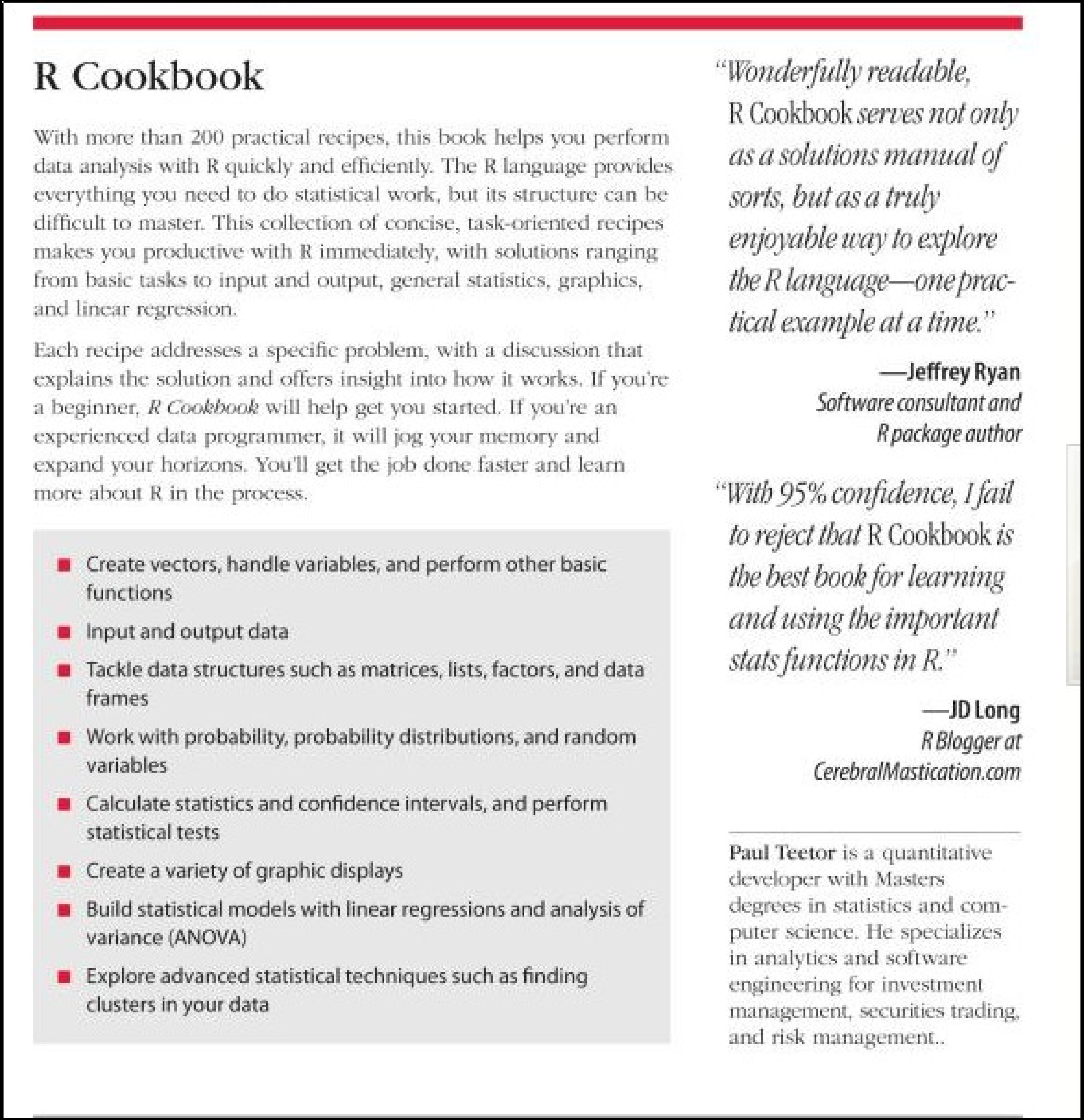
google groups
http://groups.google.com/group/segue-r

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

SEGUE

R Cookbook...





R Cookbook

With more than 200 practical recipes, this book helps you perform data analysis with R quickly and efficiently. The R language provides everything you need to do statistical work, but its structure can be difficult to master. This collection of concise, task-oriented recipes makes you productive with R immediately, with solutions ranging from basic tasks to input and output, general statistics, graphics, and linear regression.

Each recipe addresses a specific problem, with a discussion that explains the solution and offers insight into how it works. If you're a beginner, *R Cookbook* will help get you started. If you're an experienced data programmer, it will jog your memory and expand your horizons. You'll get the job done faster and learn more about R in the process.

- Create vectors, handle variables, and perform other basic functions
- Input and output data
- Tackle data structures such as matrices, lists, factors, and data frames
- Work with probability, probability distributions, and random variables
- Calculate statistics and confidence intervals, and perform statistical tests
- Create a variety of graphic displays
- Build statistical models with linear regressions and analysis of variance (ANOVA)
- Explore advanced statistical techniques such as finding clusters in your data

"Wonderfully readable, R Cookbook *serves not only as a solutions manual of sorts, but as a truly enjoyable way to explore the R language—one practical example at a time."*

—Jeffrey Ryan
Software consultant and R package author

"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
R Blogger at CerebralMastication.com

Paul Teetor is a quantitative developer with Masters degrees in statistics and computer science. He specializes in analytics and software engineering for investment management, securities trading, and risk management..

http://oreilly.com/store/dd-jpn.html