R for SysAdmins

Rigorous System Administration

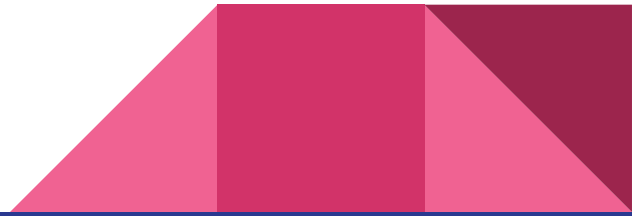
Presenter Bio

Kwan Lowe

Senior Engineer, Cloud Orchestration at Royal Caribbean Cruises

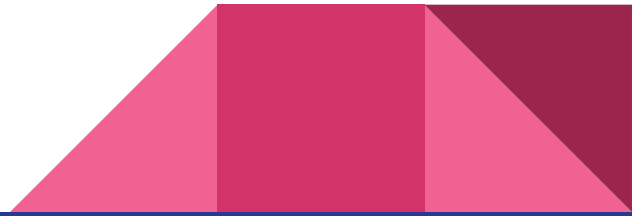
kwan@digitalhermit.com

(My first Flux presentation almost 15 years ago was on Mathematics Applications for Linux and covered R, MuPad, Octave and similar tools.)



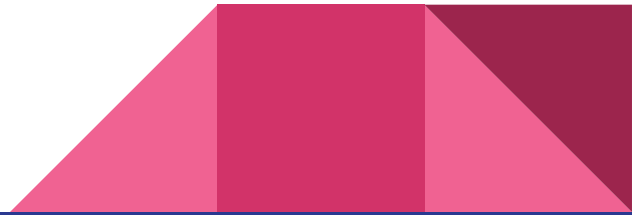
R

- Statistics language based on S
- Built for working with data (CSV, XLSX, HTML tables, etc.)
- Not a replacement for traditional scripting but "right tool" for many jobs
- GNU Licensed (Available with a "yum install R")



Long Live the SysAdmin

- Once were SysOps, Operators, BOFHs
- Roles included email administration, new user creation, filesystem maintenance
- Now SysAdmin may be responsible for scripting, full development, application management (Oracle, MySQL, Email), Security, Infrastructure (VMWare, OpenStack, AWS), growth projections, budgets, etc..



Rigorous System Administration

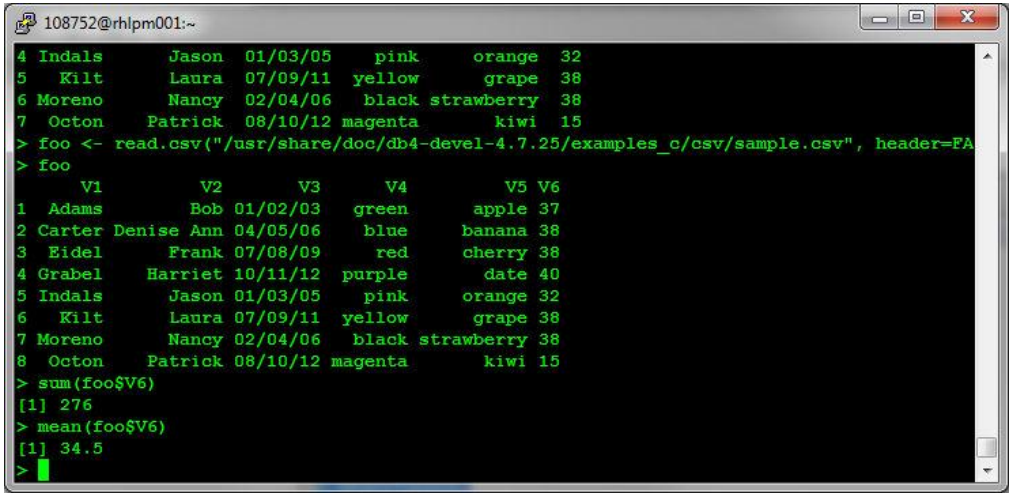
- "Proceed from knowledge."
- Years of experience can lead to "rules of thumb"
- "Rule of thumb" leads to sub-optimization and false results
- False results lead to downtime.
- Downtime leads to suffering (midnight calls).

Why R?

- Easy manipulation of data
- Scriptable, automatable
- Easy metrics (Bayesian, outliers, anomaly detection)
- Easy input/output from/to PHB-friendly formats (Excel, PDF reports)
- Language translates well to Hadoop, GPU computing

CSV, fixed column, etc.

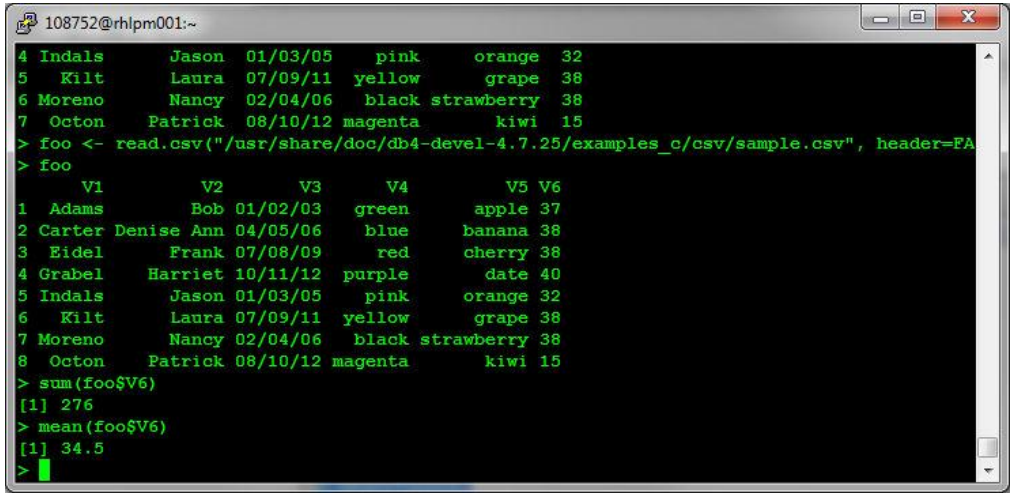
sum(foo$V6)



CSV, fixed column, etc.

But I can do that quickly in awk!!

awk '{ SUM += $6} END { print SUM}'



Excel? No problem.

```
library(xlsx)
input <- read.xlsx("myfile.xlsx"), sheet=1, header=TRUE)


output <- select(aws_volumes, VolumeId, Size, VolumeType, CreateTime)
write.xlsx(output, file = "awsVolumes.xlsx")
```

JSON? Of course...

Example from the RJSON package

```
sample_json <- '
{
"breakfast" : [ "milk", "fruit loops", "juice" ],
"lunch" : [ "left over sushi" ]
}
'

parser <- newJSONParser()

parser$addData( sample_json )
food <- parser$getObject()
print( food )

#This is equivalent to using FromJSON( sample_json )
#However, sample_json can be split into several parts:
```

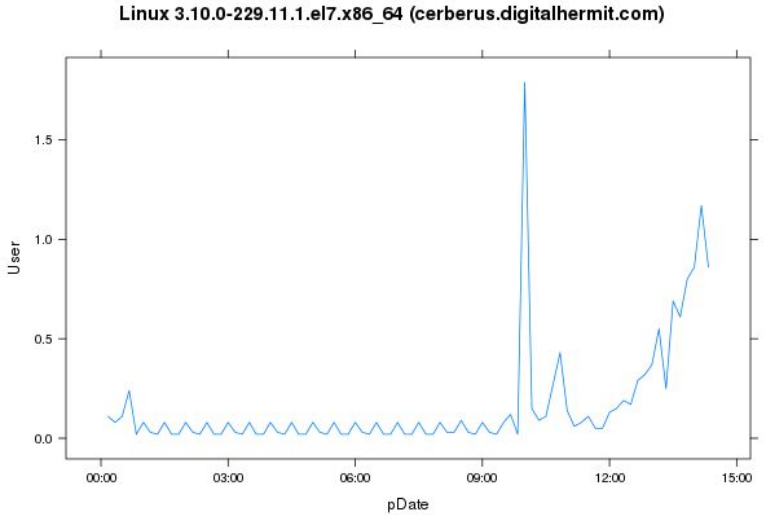
Dealing with sysstat output

sar output is messy...

```
[108752@rhldakl1069 R]$ sar

Linux 2.6.32-504.23.4.el6.x86_64 (rhldakl1069.na.rccl.com)      09/15/2015      _x86_64_(2
CPU)

12:00:01 AM     CPU      %user     %nice   %system   %iowait    %steal     %idle

12:10:02 AM     all       2.23      0.00      0.61      0.10      0.00     97.06
```

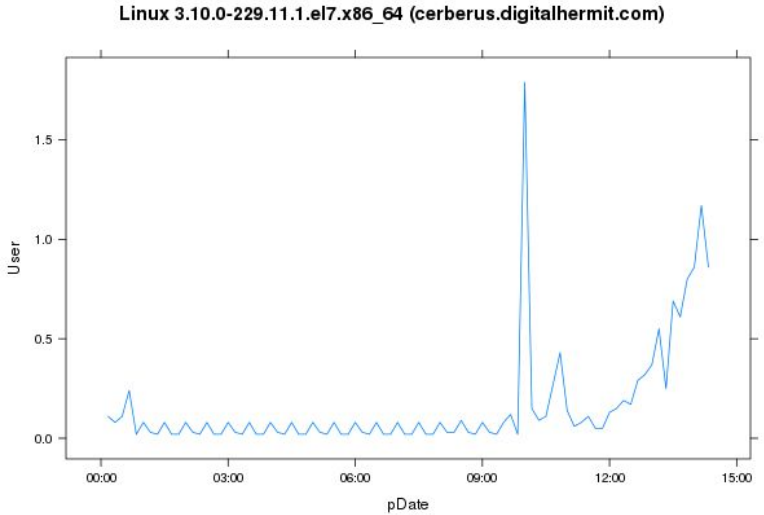
Dealing with sysstat output

sar to graph in fifteen lines of code



Dealing with sysstat output

sar to graph in fifteen lines of code



Proceed from Data

Scene in Pitch Black...

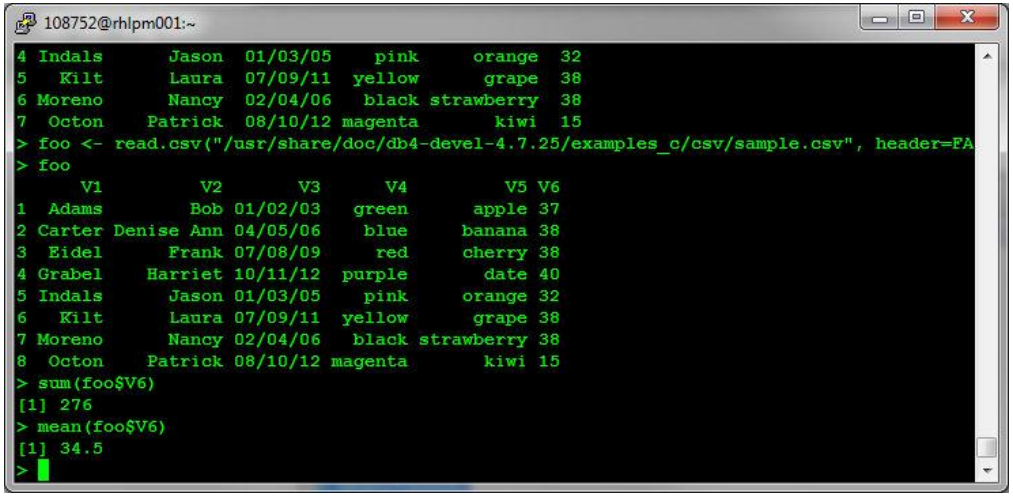
Proceed from Data

Without rigorous analysis, data can deceive…

- Experienced engineers may fall for deceptive data (cost/payoff matrix favors this approach)
- Why? Code can be hard.
- R reduces the cost of analysis.

CSV, fixed column, etc.

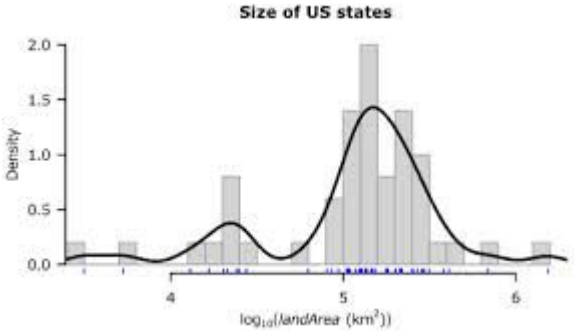
sum(foo$V6)



Proceed from Data

Quantile(and histograms) in a single line…

Here, using it to determine sizings for a template based on actual data.




```
print (quantile(mydata[,"size"]))
## END

The above will quartile the inputs to return:

    0%    25%    50%    75%   100%
   490   2024   7872  15951 123058
```

The OODA Loop

Analytics is a strategy and a tactic.

- How large should we size our virtual machine images?
- How much growth do we expect in the next year?
- What are the most common causes of outages?

Resources

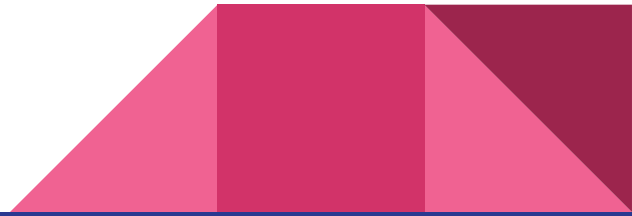
- https://cran.r-project.org/
- https://www.rstudio.com/
- https://www.coursera.org/specializations/jhudatascience
- https://github.com/twitter/AnomalyDetection
- http://www.rdatamining.com/big-data/r-hadoop-setup-guide
- http://www.r-tutor.com/gpu-computing
- https://cran.r-project.org/web/packages/rjson/

Resources

This deck...

Try It Yourself

- http://rhldakll069.na.rccl.com:8787/
- Install R/Rstudio on your laptop



Questions?