Software in Support of the Old Dominion University College of Continuing Education and Professional Development Big Data: Data Analysis Boot Camp

Chuck Cartledge

January 19, 2018

Contents

List of Figures

1 Introduction

A work in progress for software needed and used in the support of the Old Dominion University (ODU) College of Continuing Education and Professional Development (CEPD) Big Data: Data Analysis boot camp.

2 Discussion

Software will be needed on each virtual machine for the boot camp. This draft report contains a list of needed software, R scripts to install necessary libraries, and simple R scripts to test the installation.

3 Conclusion

After installing all the software identified in this report on their personal computers, the student will be able to replicate all boot camp activities.

A Software on each workstation

This section contains the assumptions about the operating system environment, and software load out for each work station.

1. Operating system: Windows 7
2. Database
   (a) Name: PostgresSQL
      - Version: 9.5.3
      - Source: `http://www.postgresql.org/download/windows/` and `http://www.enterprisedb.com/products-services-training/pgdownload#windows`
      - Superuser password: ODUBootcamp
      - Misc: It may be necessary to manually start the Postgres server using these commands in a terminal window:

        ```
        cd "\Program Files\PostgreSQL\9.5\bin"

        .\pg_ctl -D "c:\Program Files\PostgreSQL\9.5\data" start
        ```

   An R script to check out the installation is available (see Section B).
   (b) Name: Neo4j
      - Version: 3.1.2
      - Source: `https://neo4j.com/download/`
      - Superuser name: neo4j
      - Superuser password: ODUBootcamp

   The superuser name and password can be changed using a browser by going to this URL:
   `http://localhost:7474`
   If the Neo4j user name and password are lost or forgotten, they can be reset to "neo4j" and "neo4j" respectively by removing the file `data/dbms/auth` On a *nix installation, the full path name is:
   `/var/lib/neo4j/data/dbms/auth` An R script to check out the installation is available (see Section B).
3. Software
   (a) pgAdmin
      - Version: 1.22.1
      - Available from: `https://www.pgadmin.org/download/`
   (b) R
      - Version: 3.3.2
      - Available from: `https://cran.r-project.org/bin/windows/base/`
   (c) R Packages

- acs
- akima
- arules
- bda
- bit
- C50
- car
- caret
- choroplethr
- choroplethrMaps
- class
- cluster.datasets
- colorspace
- curl
- datasets
- DBI
- devtools
- doBy
- dplyr
- e1071
- ff
- ffbase
- Formula
- functional
- ggplot2
- gnm
- grid
- Hmisc
- igraph
- lattice
- maps
- MASS
- Matrix
- mlbench
- NbClust
- NLP
- party
- partykit
- plotly
- plotmo
- plotrix
- plyr
- plyrmr
- psych
- qdap
- qdapDictionaries
- qdapRegex
- qdapTools
- randomForest
- RColorBrewer
- RCurl
- readr
- reshape2
- rJava
- R.methodsS3
- RNeo4j
- ROCR
- rpart
- rpart.plot
- RPostgreSQL
- RWeka
- stringr
- survival
- TeachingDemos
- tm
- topicmodels
- utils
- vcd
- vcdExtra
- wordcloud
- wordcloud2
- xlsx
- XML

An install script is available to programmatically download the needed libraries (see Section B). To run the install script:

i. In RStudio set the session directory to the location of the install script.

ii. In the RStudio editor:

A. Type: `source(‘‘installLibraries.R’’)`

B. Press Enter (or Return)

C. Type: `source(‘‘installGitLibraries.R’’)`

D. Press Enter (or Return)

The script will download and install all the necessary libraries/packages from `https://cloud.r-project.org/`, and when completed will print `[1] ‘‘All packages installed.’’`.

(d) R-Studio

- Version: 0.99.903
- Available from: `https://www.rstudio.com/products/rstudio/download/`

(e) Hadoop

- Version: 2.7.4
- Available from: `http://hadoop.apache.org/releases.html`

(f) File decompression software

- For Unix derivatives:
  - Name: gzip
  - Version: 1.6 (at least)
  - Available from: `sudo apt-get install gzip`
- For Windows
  - Name: 7-Zip
  - Version: 16.04 (at least)
  - Available from: `http://www.7-zip.org/download.html`

(g) PDF text extraction software

- For Unix deriviatives
  - Name: pdftotext
  - Version: 0.41.0 (at least)
  - Available from: `sudoapt-getinstallpoppler-utils`
- For Windows
  - Name: XpdfReader
  - Version: 4.00
  - Available from: `http://www.xpdfreader.com/download.html`

The PATH environment variable should be updated to include the location of the R interpreter.

B Files

A collection of miscellaneous files mentioned in the report.

- installLibraries.R – an R script to install all necessary libraries/packages from "the cloud"
- installGitHubLibaries.R – an R script to install libraries/packages from `http://github.com`
- checkHadoop.R – an R script to test the Hadoop installation
- checkNeo4j02.R – an R script to test the Neo4j installation (see Figure 2).
- checkPostgres.R – an R script to test the Postgres installation (see Figure 1).

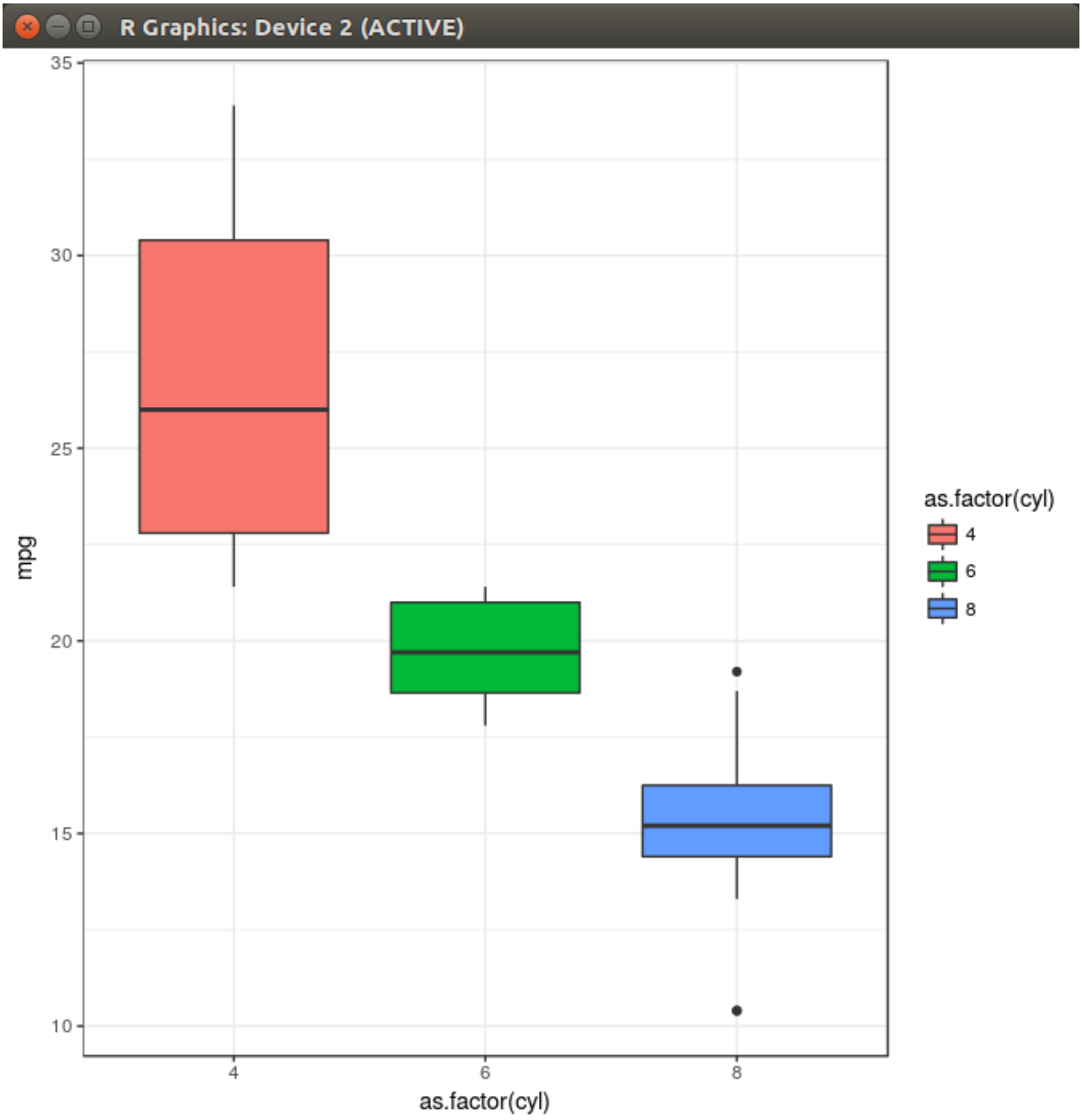

Figure 1: Image from the "checkPostgres.R" script. This image will be created (sans some of the image decorations) after successful execution of the "checkPostgres.R" script. The decorations will change based on how the script was executed.

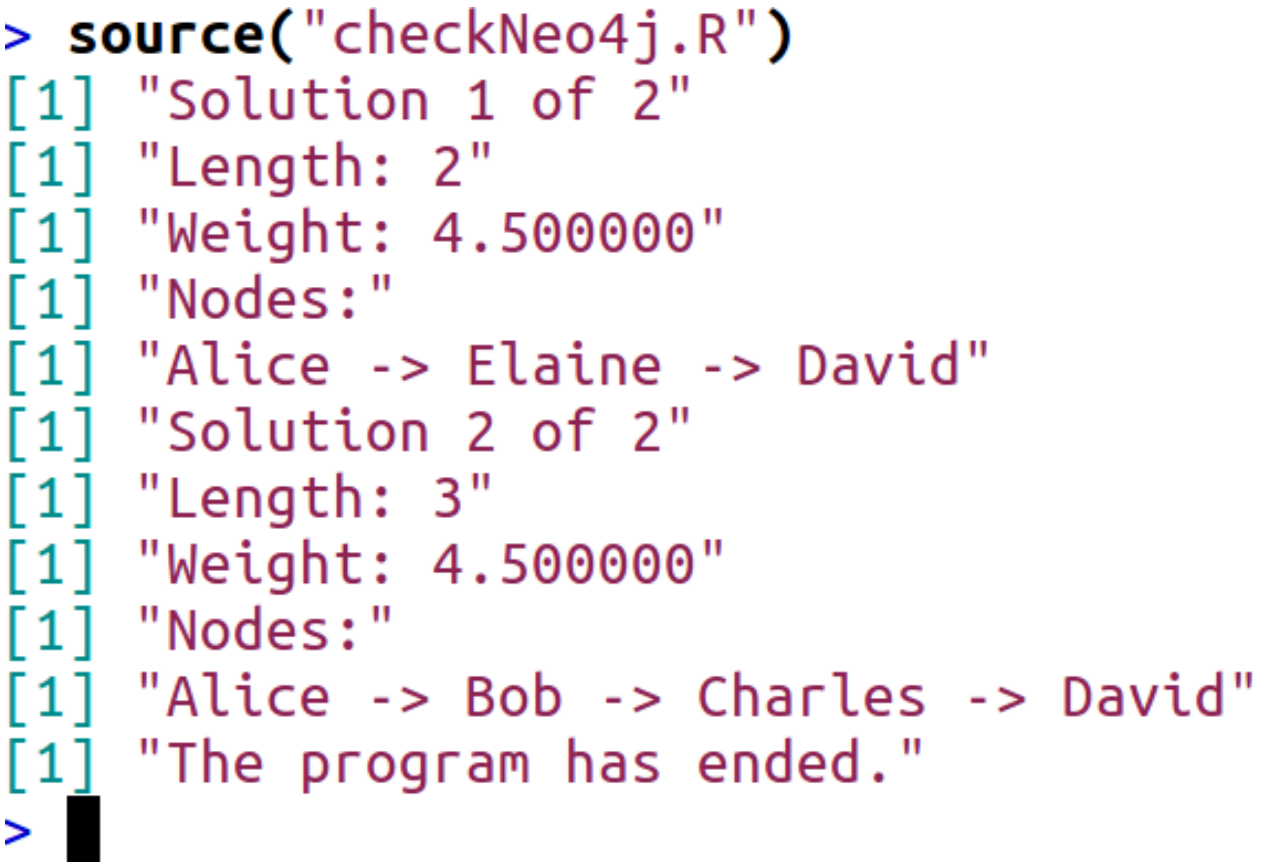

Figure 2: Image from the "checkNeo4j.R" script. This image will be created (sans some of the image decorations) after successful execution of the "checkNeo4j.R" script. The decorations will change based on how the script was executed.