**Scaling R to BIGDATA**

**Joydeep Ghosh**

R being interpretive, is inherently slow. Also it expects to be in main memory and does not intrinsically optimize for the memory hierarchy.

However there are several approaches/packages  to make R suitable for big data.

A one-stop shop is

CRAN Task View: High-Performance and Parallel Computing with R

[http://cran.r-project.org/web/views/HighPerformanceComputing.html](http://cran.r-project.org/web/views/HighPerformanceComputing.html" \t "_blank)

In general, approaches are

- writing better code

- better memory allocation

- Byte Code Compiler or R binaries

- parallelize

- interfaces with scalable libraries e.g.  [http://r-pbd.org/](http://r-pbd.org/" \t "_blank)

- translate to faster programs for core computations.

**Leading industrial product is Revolution Analytics.** [http://www.revolutionanalytics.com/](http://www.revolutionanalytics.com/" \t "_blank)

(parallelization)

described, e.g. in [http://web.warwick.ac.uk/statsdept/user2011/TalkSlides/Invited/Edlefsen-Scalable\_Data\_Analysis.pdf](http://web.warwick.ac.uk/statsdept/user2011/TalkSlides/Invited/Edlefsen-Scalable_Data_Analysis.pdf" \t "_blank)

It does provide have a free, single workstation version of R that's multi-core enabled: [http://www.revolutionanalytics.com/academic-and-public-service-programs](http://www.revolutionanalytics.com/academic-and-public-service-programs" \t "_blank)

Their offerings include (from <http://revolutionanalytics.com/news-events/revolution-analytics-introduces-revolution-r-open-and-revolution-r-plus>, 10/15/14):

1. ParallelR, a toolkit for writing parallel and distributed algorithms in the R language;
2. RHadoop, a framework for writing MapReduce algorithms in R for computation in a Hadoop cluster; and
3. DeployR Open, a Web service for R that provides for secure integration of real-time R calculations into third-party applications.

Book: Ch 24 of "**R in a Nutshell**" talks about optimizing R.

Ch 26 is on supporting R on Hadoop.

Some additional resources that might help (Hadley Wickham is quite amazing):  
[http://adv-r.had.co.nz/memory.html](http://adv-r.had.co.nz/memory.html" \t "_blank)  
[http://stackoverflow.com/questions/1358003/tricks-to-manage-the-available-memory-in-an-r-session](http://stackoverflow.com/questions/1358003/tricks-to-manage-the-available-memory-in-an-r-session" \t "_blank)

**Python and Big Data**

Python support on Hadoop/Spark has started appearing.

Conda: Anaconda Cluster announced 10/14

<http://continuum.io/anaconda-cluster>

PySpark is a python API for Spark

<http://spark.apache.org/docs/latest/api/python/>

includes several ML libraries