Scalable Data Analysis in *R*

Lee E. Edlefsen, Ph.D.[1*]

1. Chief Scientist, Revolution Analytics, Inc.
*Contact author: lee@revolutionanalytics.com

Keywords: Scalable analysis, large data, parallel computing, distributed computing, high performance computing

For the past several decades the rising tide of technology -- especially the increasing speed of single processors -- has allowed the same data analysis code to run faster and on bigger data sets. That happy era is ending. The size of data sets is increasing much more rapidly than the speed of single cores, of I/O, and of RAM. To deal with this, we need software that can use multiple cores, multiple hard drives, and multiple computers.

That is, we need scalable data analysis software. It needs to scale from small data sets to huge ones, from using one core and one hard drive on one computer to using many cores and many hard drives on many computers, and from using local hardware to using remote clouds.

R is the ideal platform for scalable data analysis software. It is easy to add new functionality in the *R* environment, and easy to integrate it into existing functionality. *R* is also powerful, flexible and forgiving.

I will discuss the approach to scalability we have taken at Revolution Analytics with our package **RevoScaleR.** A key part of this approach is to efficiently operate on "chunks" of data -- sets of rows of data for selected columns. I will discuss this approach from the point of view of:

- Storing data on disk
- Importing data from other sources
- Reading and writing of chunks of data
- Handling data in memory
- Using multiple cores on single computers
- Using multiple computers
- Automatically parallelizing "external memory" algorithms