Big Data Analytics with R and Hadoop

Jamie F Olson[1,⋆]

1. Microsoft Algorithms & Data Science
⋆Contact author: jamieo@microsoft.com

Keywords: Hadoop, R

R is a powerful and flexible language for statistical programming and exploratory data analysis, but it has long been hampered by its performance with large datasets. Despite significant recent performance improvements, *R* is limited by design to data that fits in the RAM of a single workstation or server. This fundamental limitation can be overcome in a variety of ways including both distributed computing and external memory algorithms, both of which can be leveraged using Revolution R Enterprise(RRE) ScaleR. I will present an overview of the RRE reference architectures for Hadoop as well as a overview of the ScaleR features and application programming interface. In addition to RRE, I will provide a brief overview of some additional tools for utilizing the Hadoop from *R*, including the RHadoop packages. By carefully combining *R* and its many packages with the powerful Hadoop distributed computing platform we can overcome *R*'s memory limitations and still leverage *R*'s incredible productivity.