Running Hadoop and Spark from R Using Docker Containers

E. James Harner[1*], **Mark Lilback**[1]

1. Department of Statistics, West Virginia University
*Contact author: jharner@stat.wvu.edu

Keywords: Hadoop and Spark, Mesos, Docker containers, R, Cloud computing

There are numerous obstacles in accessing HDFS/Hadoop and Spark from *R*. Scripts and packages must be distributed and possibly compiled, which is difficult for those not intimately familiar with the command line. rc^2 (*R* Cloud Computing) is an experimental environment roughly based on the Berkeley Data Analytics Stack (BDAS). A prototype OS X/ iOS client is also being developed.

The server side of rc^2 is built on Mesos, a distributed system kernel. Mesos allows the launching of tasks containing Docker images and in the future the more flexible, more secure Rocker containers are likely to be supported. Initially, containers will be used to run multiple instances of *R*. These instances support regular *R* sessions, but can also access the HDFS/Hadoop and Spark ecosystems. Existing packages, including **RHadoop**, **RHIPE**, and **SparkR**, provide interfaces to Hadoop/ Spark. The cluster frameworks, such as Spark and MapReduce, can also be run in containers in a single VM on a laptop or on a cloud provider such as AWS or Azure.