Running Hadoop and Spark from R Using Docker Containers

E. James Harner\textsuperscript{1,*}, Mark Lilback\textsuperscript{1}

\textsuperscript{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 \texttt{R}. Scripts and packages must be distributed and possibly compiled, which is difficult for those not intimately familiar with the command line. \texttt{rc2 (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 \texttt{rc2} is built on \texttt{Mesos}, a distributed system kernel. Mesos allows the launching of tasks containing \texttt{Docker} images and in the future the more flexible, more secure \texttt{Rocker} containers are likely to be supported. Initially, containers will be used to run multiple instances of \texttt{R}. These instances support regular \texttt{R} sessions, but can also access the HDFS/Hadoop and Spark ecosystems. Existing packages, including \texttt{RHadoop}, \texttt{RHIPE}, and \texttt{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.